MCA 205 DATA SCIENCE ESSENTIALS

Unit-I

Introduction: What is Data Science? - Big Data and Data Science, Statistical Inference - Populations and samples - Statistical modeling, probability distributions, fitting a model - Intro to R Language.

UNIT-II

Exploratory Data Analysis and the Data Science Process - Basic tools (plots, graphs and summary statistics) of EDA - Philosophy of EDA - The Data Science Process - Case Study: RealDirect (online real estate firm)

UNIT-III

Feature Generation and Feature Selection (Extracting Meaning From Data) - Motivating application: user (customer) retention - Feature Generation (brainstorming, role of domain expertise, and place for Imaginationl - Feature Selection algorithms - Filters; Wrappers; Decision Trees; Random Forests

UNIT-IV

Data Visualization - Basic principles, ideas and tools for data visualization 3 - Examples of inspiring (industry) projects - Exercise: create your own visualization of a complex dataset

UNIT-V

Data Science and Ethical Issues - Discussions on privacy, security, ethics - A look back at Data Science - Next-generation data scientists

Text Book:

1. Cathy O'Neil and Rachel Schutt. Doing Data Science,' Straight Talk From The Frontline. O'Reilly. 2014.

REFERENCES BOOKS:

1. Jure Leskovek, Anand Rajaraman and Jeffrey Ullman. Mining of Massive Datasets. v2.1, Cambridge University Press. 2014. (free online)
2. Kevin P. Murphy. Machine Learning: A Probabilistic Perspective. ISBN 0262018020.2013.
3. Foster Provost and Tom Fawcett. Data Science for Business: What You Need to Know about Data Mining and Data-analytic Thinking. ISBN 1449361323. 2013.
4. Trevor Hastie, Robert Tibshirani and Jerome Friedman. Elements of Statistical Learning, Second Edition. ISBN 0387952845. 2009. (free online)

SAITEJA XEROX

Scanned with CamScanner

UNIT – I

Introduction

What is Data Science?

Data science, also known as **data-driven science**, is an interdisciplinary field of scientific methods, processes, and systems to extract knowledge or insights from data in various forms, either structured or unstructured, similar to data mining.

Data science is a "concept to unify statistics, data analysis and their related methods" in order to "understand and analyze actual phenomena" with data. It employs techniques and theories drawn from many fields within the broad areas of mathematics, statistics, information science, and computer science, in particular from the subdomains of machine learning, classification, cluster analysis, data mining, databases, and visualization.

Turing award winner Jim Gray imagined data science as a "fourth paradigm" of science (empirical, theoretical, computational and now data-driven) and asserted that "everything about science is changing because of the impact of information technology" and the data deluge.

DEFINITION of 'Data Science'

A field of Big Data which seeks to provide meaningful information from large amounts of complex data. Data Science combines different fields of work in statistics and computation in order to interpret data for the purpose of decision making.

Data science is the study of where information comes from, what it represents and how it can be turned into a valuable resource in the creation of business and IT strategies. Mining large amounts of structured and unstructured data to identify patterns can help an organization rein in costs, increase efficiencies, recognize new market opportunities and increase the organization's competitive advantage.

History

The term "data science" (originally used interchangeably with "datalogy") has existed for over thirty years and was used initially as a substitute for computer science by Peter Naur in 1960. In 1974, Naur published *Concise Survey of Computer Methods*, which freely used the term data science in its survey of the contemporary data processing methods that are used in a wide range of applications.

In 1996, members of the International Federation of Classification Societies (IFCS) met in Kobe for their biennial conference. Here, for the first time, the term data science is included in the title of the conference

SAITEJA XEROX

Scanned with CamScanner

("Data Science, classification, and related methods"),after the term was introduced in a roundtable discussion by Chikio Hayashi.

Data science process flowchart from "Doing Data Science", Cathy O'Neil and Rachel Schutt, 2013

Big Data and Data science

The terms Big Data and Data Science are associated with large volumes of data characterizing the new technological era. In particular, with the collection, analysis and, as an ultimate objective, extraction value of such data to aid in decision making.

Both are closely related conceptually, but in no case are synonymous terms.

What is Big Data?

The concept relates to the efficient collection of a large volume of heterogeneous data (not stored in a traditional database) that can be structured, semi-structured or unstructured, to the storage and analysis in a short time, most sometimes in real time.

Although it is clear in general terms, it is a novel concept that encompasses a much broader scope of strictly technological -the term was introduced in the Oxford Dictionary in 2013-, becoming a buzzword.

In the absence of a universal definition, since there is no agreement on what is the "big data", it has been proposed very different definitions that, under the interest in everything related to Big Data, they have much less exhausted all the possibilities.

In general terms, we may agree that "Big Data" usually refers to large data, basically the scale of terabytes and petabytes (a petabyte is a million gigabytes), and its potential to deepen in our understanding of the phenomena that arouse our interest.

From the "physical and biological systems of human social and economic behaviour", as the UC Berkeley Scool of Information notes on its website, to scientific objectives, business, concerning public administration or, of course, to any other susceptible area analysis.

SAITEJA XEROX

Scanned with CamScanner

No doubt that the required analyzes to process the huge amount of data require technical resources and specific IT and algorithms that brings Data Science, a discipline that is in full swing, and it continues to grow under the umbrella of Big Data.

Just as is true in reverse, because thanks to data science and new technologies, characterized by high efficiency, Big Data transcends the phenomenon of big data to reach a higher level.

Here, we find the connection between the two concepts, and its differentiation too. Through its sum, we obtain unimagined synergies. For the first time in history it is feasible to extract value thereof at low cost, making it available to organizations in the private, public or scientific sector as never it had been before.

The practical functionality of Big Data we get it, therefore, working with data in order to open infinite possibilities of advancement for competitive advantages, according to the emerging needs. And, as it has been noted, this value only gives us Data Science.

Data Science, The Big Data Key

Based on the concept of Big Data we have arrived at Data Science. A holistic approach would allow us to say that this is part of that universe, but no longer a different concept.

The figure of data scientist is a key in Data Science. And as its name suggests, is the scientific data who applies relying on ideas, mathematical, statistical and computer tools; with them, this works to make intelligent analysis of large data.

Always aligning the goals of the organization or the scientific team concerned by the use of technology to find solutions, forecasting, providing information in real-time, accessible through different channels, and by an easy visualization of the results.

In the current context, data science it is a driver of Big Data, giving it with an unprecedented potential. As a master key that is, it helps us to take advantage of Big Data in a versatile way, and despite its breadth and casuistry concept, its ultimate goal is to move forward in key forward.

Both towards establishing comparative advantages as gaining knowledge in general within the new framework of analysis provided by the big data, presented, not in the typical columns and rows, but hundreds of billions of rows and millions of variables.

It supposes, therefore, a paradigm shift from the traditional data analysis, representing a turning point regarding databases and traditional BI. Still, hybrid systems remain an interesting alternative.

That is reflected in the report published by McKinsey Global Institute (MGI) in June 2011, when it defines Big Data as "data sets whose size is beyond the ability of capture, storage, management and analysis tools databases".

SAITEJA XEROX

Scanned with CamScanner

Despite its differences, in short, both concepts are closely related and interdependent. That-is, at least in the present context we live, dominated by the digital era in which big data and its use are inextricably linked when Big Data and Data Science is all about.

Statistical inference

Statistical inference is the process of deducing properties of an underlying probability distribution by analysis of data. Inferential statistical analysis infers properties about a population: this includes testing hypotheses and deriving estimates. The population is assumed to be larger than the observed data set; in other words, the observed data is assumed to be sampled from a larger population.

Inferential statistics can be contrasted with descriptive statistics. Descriptive statistics is solely concerned with properties of the observed data, and does not assume that the data came from a larger population.

Introduction

Statistical inference makes propositions about a population, using data drawn from the population with some form of sampling. Given a hypothesis about a population, for which we wish to draw inferences, statistical inference consists of (first) selecting a statistical model of the process that generates the data and (second) deducing propositions from the model.

Konishi& Kitagawa state, "The majority of the problems in statistical inference can be considered to be problems related to statistical modeling". Relatedly, Sir David Cox has said, "How [the] translation from subject-matter problem to statistical model is done is often the most critical part of an analysis".

The conclusion of a **statistical inference** is a statistical proposition. Some common forms of statistical proposition are the following:

- a point estimate, i.e. a particular value that best approximates some parameter of interest;
- an interval estimate, e.g. a confidence interval (or set estimate), i.e. an interval constructed using a dataset drawn from a population so that, under repeated sampling of such datasets, such intervals would contain the true parameter value with the probability at the stated confidence level;
- a credible interval, i.e. a set of values containing, for example, 95% of posterior belief;
- rejection of a hypothesis;
- clustering or classification of data points into groups.

Scanned with CamScanner

Models and assumptions

Any statistical inference requires some assumptions. A **statistical model** is a set of assumptions concerning the generation of the observed data and similar data. Descriptions of statistical models usually emphasize the role of population quantities of interest, about which we wish to draw inference. Descriptive statistics are typically used as a preliminary step before more formal inferences are drawn.

Degree of models/assumptions

Statisticians distinguish between three levels of modeling assumptions;

- **Fully parametric**: The probability distributions describing the data-generation process are assumed to be fully described by a family of probability distributions involving only a finite number of unknown parameters. For example, one may assume that the distribution of population values is truly Normal, with unknown mean and variance, and that datasets are generated by 'simple' random sampling. The family of generalized linear models is a widely used and flexible class of parametric models.
- **Non-parametric**: The assumptions made about the process generating the data are much less than in parametric statistics and may be minimal. For example, every continuous probability distribution has a median, which may be estimated using the sample median or the Hodges–Lehmann–Sen estimator, which has good properties when the data arise from simple random sampling.
- **Semi-parametric**: This term typically implies assumptions 'in between' fully and non-parametric approaches. For example, one may assume that a population distribution has a finite mean. Furthermore, one may assume that the mean response level in the population depends in a truly linear manner on some covariate (a parametric assumption) but not make any parametric assumption describing the variance around that mean (i.e. about the presence or possible form of any heteroscedasticity). More generally, semi-parametric models can often be separated into 'structural' and 'random variation' components. One component is treated parametrically and the other non-parametrically. The well-known Cox model is a set of semi-parametric assumptions.

Importance of valid models/assumptions

Whatever level of assumption is made, correctly calibrated inference in general requires these assumptions to be correct; i.e. that the data-generating mechanisms really have been correctly specified.

SAITEJA XEROX

Scanned with CamScanner

Incorrect assumptions of 'simple' random sampling can invalidate statistical inference. More complex semi- and fully parametric assumptions are also cause for concern. For example, incorrectly assuming the Cox model can in some cases lead to faulty conclusions. Incorrect assumptions of Normality in the population also invalidates some forms of regression-based inference. The use of any parametric model is viewed skeptically by most experts in sampling human populations: "most sampling statisticians, when they deal with confidence intervals at all, limit themselves to statements about [estimators] based on very large samples, where the central limit theorem ensures that these [estimators] will have distributions that are nearly normal." In particular, a normal distribution "would be a totally unrealistic and catastrophically unwise assumption to make if we were dealing with any kind of economic population." Here, the central limit theorem states that the distribution of the sample mean "for very large samples" is approximately normally distributed, if the distribution is not heavy tailed.

Paradigms for inference

Different schools of statistical inference have become established. These schools—or "paradigms"—are not mutually exclusive, and methods that work well under one paradigm often have attractive interpretations under other paradigms.

Bandyopadhyay& Forster describe four paradigms: "(i) classical statistics or error statistics, (ii) Bayesian statistics, (iii) likelihood-based statistics, and (iv) theAkaikean-Information Criterion-based statistics".

Populations and Samples

Population

A **population** is a set of similar items or events which is of interest for some question or experiment. A statistical population can be a group of actually existing objects (e.g. the set of all stars within the Milky Way galaxy) or a hypothetical and potentially infinite group of objects conceived as a generalization from experience (e.g. the set of all possible hands in a game of poker). A common aim of statistical analysis is to produce information about some chosen population.

In statistical inference, a subset of the population (a statistical sample) is chosen to represent the population in a statistical analysis. If a sample is chosen properly, characteristics of the entire population that the sample is drawn from can be estimated from corresponding characteristics of the sample

Subpopulation

A subconcept of a population that shares one or more additional properties is called a **subpopulation**. For example, if the population is all Egyptian people, a subpopulation is all Egyptian males; if the population

SAITEJA XEROX

Scanned with CamScanner

is all pharmacies in the world, a subpopulation is all pharmacies in Egypt. By contrast, a sample is a subset of a population that is not chosen to share any additional property.

Descriptive statistics may yield different results for different subpopulations. For instance, a particular medicine may have different effects on different subpopulations, and these effects may be obscured or dismissed if such special subpopulations are not identified and examined in isolation.

Similarly, one can often estimate parameters more accurately if one separates out subpopulations: the distribution of heights among people is better modeled by considering men and women as separate subpopulations, for instance.

Populations consisting of subpopulations can be modeled by mixture models, which combine the distributions within subpopulations into an overall population distribution. Even if subpopulations are well-modeled by given simple models, the overall population may be poorly fit by a given simple model – poor fit may be evidence for existence of subpopulations. For example, given two equal subpopulations, both normally distributed, if they have the same standard deviation and different means, the overall distribution will exhibit low kurtosis relative to a single normal distribution – the means of the subpopulations fall on the shoulders of the overall distribution. If sufficiently separated, these form a bimodal distribution, otherwise it simply has a wide peak. Further, it will exhibit overdispersion relative to a single normal distribution with the given variation. Alternatively, given two subpopulations with the same mean and different standard deviations, the overall population will exhibit high kurtosis, with a sharper peak and heavier tails (and correspondingly shallower shoulders) than a single distribution.

Sample

A **data sample** is a set of data collected and/or selected from a statistical population by a defined procedure. The elements of a sample are known as **sample points, sampling units** or observations.

Typically, the population is very large, making a census or a complete enumeration of all the values in the population either impractical or impossible. The sample usually represents a subset of manageable size. Samples are collected and statistics are calculated from the samples, so that one can make inferences or extrapolations from the sample to the population.

The data sample may be drawn from a population *without replacement* (i.e. no element can be selected more than once in the same sample), in which case it is a subset of a population; or *with replacement* (i.e. an element may appear multiple times in the one sample), in which case it is a multisubset.

SAI TEJA XEROX

Scanned with CamScanner

Kinds of samples

A **complete sample** is a set of objects from a parent population that includes ALL such objects that satisfy a set of well-defined selection criteria. For example, a complete sample of Australian men taller than 2m would consist of a list of every Australian male taller than 2m. But it wouldn't include German males, or tall Australian females, or people shorter than 2m. So to compile such a complete sample requires a complete list of the parent population, including data on height, gender, and nationality for each member of that parent population. In the case of human populations, such a complete list is unlikely to exist (the human population being in the billions). But, such complete samples are often available in other disciplines such as the set of players in a major sports league, the birth dates of the members of a parliament, or a complete magnitude-limited list of astronomical objects.

An **unbiased (representative) sample** is a set of objects chosen from a complete sample using a selection process that does not depend on the properties of the objects. For example, an unbiased sample of Australian men taller than 2m might consist of a randomly sampled subset of 1% of Australian males taller than 2m. But one chosen from the electoral register might not be unbiased since, for example, males aged under 18 will not be on the electoral register. In an astronomical context, an unbiased sample might consist of that fraction of a complete sample for which data are available, provided the data availability is not biased by individual source properties.

The best way to avoid a biased or unrepresentative sample is to select a random sample, also known as a probability sample. A random sample is defined as a sample where each individual member of the population has a known, non-zero chance of being selected as part of the sample. Several types of random samples are simple random samples, systematic samples, stratified random samples, and cluster random samples.

A sample that is not random is called a non-random sample or a non-probability sampling. Some examples of nonrandom samples are convenience samples, judgment samples, purposive samples, quota samples, snowball samples, and quadrature nodes in quasi-Monte Carlo methods.

Statistic samples have multiple uses. They can be used in many situations.

Population vs Sample

The main difference between a population and sample has to do with how observations are assigned to the data set.

- A **population** includes all of the elements from a set of data.
- A **sample** consists one or more observations drawn from the population.

SAI TEJA XEROX

Scanned with CamScanner

Depending on the sampling method, a sample can have fewer observations than the population, the same number of observations, or more observations. More than one sample can be derived from the same population.

What is Simple Random Sampling?

A **sampling method** is a procedure for selecting sample elements from a population. **Simple random samplingrefers** to a sampling method that has the following properties.

- The population consists of N objects.
- The sample consists of n objects.
- All possible samples of n objects are equally likely to occur.

An important benefit of simple random sampling is that it allows researchers to use statistical methods to analyze sample results. For example, given a simple random sample, researchers can use statistical methods to define aconfidence interval around a sample mean. Statistical analysis is not appropriate when non-random sampling methods are used.

There are many ways to obtain a simple random sample. One way would be the lottery method. Each of the Npopulation members is assigned a unique number. The numbers are placed in a bowl and thoroughly mixed. Then, a blind-folded researcher selects n numbers. Population members having the selected numbers are included in the sample.

Sampling With Replacement and Without Replacement

Suppose we use the lottery method described above to select a simple random sample. After we pick a number from the bowl, we can put the number aside or we can put it back into the bowl. If we put the number back in the bowl, it may be selected more than once; if we put it aside, it can selected only one time.

When a population element can be selected more than one time, we are **sampling with replacement.** When a population element can be selected only one time, we are **sampling without replacement.**

Statistical modeling

A **statistical model** is a class of mathematical model, which embodies a set of assumptions concerning the generation of some sample data, and similar data from a larger population. A statistical model represents, often in considerably idealized form, the data-generating process.

The assumptions embodied by a statistical model describe a set of probability distributions, some of which are assumed to adequately approximate the distribution from which a particular data set is sampled.

SAITEJA XEROX

Scanned with CamScanner

The probability distributions inherent in statistical models are what distinguishes statistical models from other, non-statistical, mathematical models.

A statistical model is usually specified by mathematical equations that relate one or more random variables and possibly other non-random variables. As such, a statistical model is "a formal representation of a theory" (Herman Adèr quoting Kenneth Bollen).

All statistical hypothesis tests and all statistical estimators are derived from statistical models. More generally, statistical models are part of the foundation of statistical inference.

Formal definition

In mathematical terms, a statistical model is usually thought of as a pair $(S, \mathcal{P})$, where S is the set of possible observations, i.e. the sample space, and $\mathcal{P}$ is a set of probability distributions on S.

The intuition behind this definition is as follows. It is assumed that there is a "true" probability distribution induced by the process that generates the observed data. We choose $\mathcal{P}$ to represent a set (of distributions) which contains a distribution that adequately approximates the true distribution. Note that we do not require that $\mathcal{P}$ contains the true distribution, and in practice that is rarely the case. Indeed, as Burnham & Anderson state, "A model is a simplification or approximation of reality and hence will not reflect all of reality" —whence the saying "all models are wrong".

The set $\mathcal{P}$ is almost always parameterized: $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$. The set Θ defines the parameters of the model. A parameterization is generally required to have distinct parameter values give rise to distinct distributions, i.e. $P_{\theta_1} = P_{\theta_2} \Rightarrow \theta_1 = \theta_2$ must hold (in other words, it must be injective). A parameterization that meets the requirement is said to be identifiable.

An example

Suppose that we have a population of school children, with the ages of the children distributed uniformly, in the population. The height of a child will be stochastically related to the age: e.g. when we know that a child is of age 7, this influences the chance of the child being 5 feet tall. We could formalize that relationship in a linear regression model, like this: $\text{height}_i = b_0 + b_1\text{age}_i + \varepsilon_i$, where b_0 is the intercept, b_1 is a parameter that age is multiplied by in obtaining a prediction of height, ε_i is the error term, and i identifies the child. This implies that height is predicted by age, with some error.

An admissible model must be consistent with all the data points. Thus, a straight line ($\text{height}_i = b_0 + b_1\text{age}_i$) cannot be the equation for a model of the data. The line cannot be the equation for a model, unless it exactly fits all the data points—i.e. all the data points lie perfectly on the line. The error term, ε_i, must be included in the equation, so that the model is consistent with all the data points.

To do statistical inference, we would first need to assume some probability distributions for the ε_i. For instance, we might assume that the ε_i distributions are i.i.d. Gaussian, with zer mean. In this instance, the model would have 3 parameters: b_0, b_1, and the variance of the Gaussian distribution.

We can formally specify the model in the form $(S, \mathcal{P})$ as follows. The sample space, S, of our model comprises the set of all possible pairs (age, height). Each possible value of $\theta = (b_0, b_1, \sigma^2)$ determines a distribution on S; denote that distribution by P_θ. If Θ is the set of all possible values of θ, then $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$. (The parameterization is identifiable, and this is easy to check.)

In this example, the model is determined by (1) specifying S and (2) making some assumptions relevant to $\mathcal{P}$. There are two assumptions: that height can be approximated by a linear function of age; that errors in the approximation are distributed as i.i.d. Gaussian. The assumptions are sufficient to specify $\mathcal{P}$—as they are required to do.

Scanned with CamScanner

Dimension of a model

Suppose that we have a statistical model $(S, \mathcal{P})$ with $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$. The model is said to be *parametric* if Θ has a finite dimension. In notation, we write that $\Theta \subseteq \mathbb{R}^k$ where k is a positive integer ($\mathbb{R}$ denotes the real numbers; other sets can be used, in principle). Here, k is called the **dimension** of the model.

As an example, if we assume that data arise from a univariate Gaussian distribution, then we are assuming that

$$\mathcal{P} = \left\{ P_{\mu,\sigma}(x) \equiv \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2} \right) : \mu \in \mathbb{R}, \sigma > 0 \right\}.$$

In this example, the dimension, k, equals 2.

As another example, suppose that the data consists of points (x, y) that we assume are distributed according to a straight line with i.i.d. Gaussian residuals (with zero mean). Then the dimension of the statistical model is 3: the intercept of the line, the slope of the line, and the variance of the distribution of the residuals. (Note that in geometry, a straight line has dimension 1.)

Although formally $\theta \in \Theta$ is a single parameter that has dimension k, it is sometimes regarded as comprising k separate parameters. For example, with the univariate Gaussian distribution, θ is a single parameter with dimension 2, but it is sometimes regarded as comprising 2 separate parameters—the mean and the standard deviation.

A statistical model is *nonparametric* if the parameter set Θ is infinite dimensional. A statistical model is *semiparametric* if it has both finite-dimensional and infinite-dimensional parameters. Formally, if k is the dimension of Θ and n is the number of samples, both semiparametric and nonparametric models have $k \to \infty$ as $n \to \infty$. If $k/n \to 0$ as $n \to \infty$ then the model is semiparametric; otherwise, the model is nonparametric.

Parametric models are by far the most commonly used statistical models. Regarding semiparametric and nonparametric models, Sir David Cox has said, "These typically involve fewer assumptions of structure and distributional form but usually contain strong assumptions about independencies".

Scanned with CamScanner

Nested models

Two statistical models are **nested** if the first model can be transformed into the second model by imposing constraints on the parameters of the first model. As an example, the set of all Gaussian distributions has, nested within it, the set of zero-mean Gaussian distributions. we constrain the mean in the set of all Gaussian distributions to get the zero-mean distributions. As a second example, the quadratic model

$$y = b_0 + b_1x + b_2x^2 + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2)$$

has, nested within it, the linear model

$$y = b_0 + b_1x + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2)$$

—we constrain the parameter b_2 to equal 0

In both those examples, the first model has a higher dimension than the second model (for the first example, the zero-mean model has dimension 1). Such is usually, but not always, the case. As a different example, the set of positive-mean Gaussian distributions, which has dimension 2, is nested within the set of all Gaussian distributions.

Comparing models

It is assumed that there is a "true" probability distribution underlying the observed data, induced by the process that generated the data. The main goal of model selection is to make statements about which elements of $\mathcal{P}$ are most likely to adequately approximate the true distribution.

Models can be compared to each other by exploratory data analysis or confirmatory data analysis. In exploratory analysis, a variety of models are formulated and an assessment is performed of how well each one describes the data. In confirmatory analysis, a previously formulated model or models are compared to the data. Common criteria for comparing models include R^2, Bayes factor, and the likelihood-ratio test together with its generalization relative likelihood.

Konishi & Kitagawa state: "The majority of the problems in statistical inference can be considered to be problems related to statistical modeling. They are typically formulated as comparisons of several statistical models." Relatedly, Sir David Cox has said, "How [the] translation from subject-matter problem to statistical model is done is often the most critical part of an analysis".

Probability distributions

A **probability distribution** is a mathematical function that, stated in simple terms, can be thought of as providing the probabilities of occurrence of different possible outcomes in an experiment. For instance, if the random variable X is used to denote the outcome of a coin toss ("the experiment"), then the probability distribution of X would take the value 0.5 for $X =$ heads, and 0.5 for $X =$ tails (assuming the coin is fair).

In more technical terms, the probability distribution is a description of a random phenomenon in terms of the probabilities of events. Examples of random phenomena can include the results of an experiment or survey. A probability distribution is defined in terms of an underlying sample space, which is the set of all possible outcomes of the random phenomenon being observed. The sample space may be the set of real numbers or a higher-dimensional vector space, or it may be a list of non-numerical values; for example, the sample space of a coin flip would be {heads, tails} .

Probability distributions are generally divided into two classes. A **discrete probability distribution** (applicable to the scenarios where the set of possible outcomes is discrete, such as a coin toss or a roll of dice) can be encoded by a discrete list of the probabilities of the outcomes, known as

Scanned with CamScanner

a probability mass function. On the other hand, a **continuous probability distribution** (applicable to the scenarios where the set of possible outcomes can take on values in a continuous range (e.g. real numbers), such as the temperature on a given day) is typically described by probability density functions (with the probability of any individual outcome actually being 0). The normal distribution is a commonly encountered continuous probability distribution. More complex experiments, such as those involving stochastic processes defined in continuous time, may demand the use of more general probability measures.

A probability distribution whose sample space is the set of real numbers is called univariate, while a distribution whose sample space is a vector space is called multivariate. A univariate distribution gives the probabilities of a single random variable taking on various alternative values; a multivariate distribution (a joint probability distribution) gives the probabilities of a random vector—a list of two or more random variables—taking on various combinations of values. Important and commonly encountered univariate probability distributions include the binomial distribution, the hypergeometric distribution, and the normal distribution. The multivariate normal distribution is a commonly encountered multivariate distribution.

Introduction

To define probability distributions for the simplest cases, one needs to distinguish between **discrete** and **continuous** random variables In the discrete case, it is sufficient to specify a probability mass function p assigning a probability to each possible outcome: for exampl when throwing a fair dice, each of the six values 1 to 6 has the probability 1/6. The probability of an event is then defined to be the sum of the probabilities of the outcomes that satisfy the event; for example, the probability of the event "the dice rolls an even value" is

$$p(2) + p(4) + p(6) = 1/6 + 1/6 + 1/6 = 1/2.$$

In contrast, when a random variable takes values from a continuum then typically, any individual outcome has probability zero and only events that include infinitely many outcomes, such as intervals, can have positive probability. For example, the probability that a given object weighs *exactly* 500 g is zero, because the probability of measuring exactly 500 g tends to zero as the accuracy of our measuring instruments increases. Nevertheless, in quality control one might demand that the probability of a "500 g" package containing between 490 g and 510 g should be no less than 98%, and this demand is less sensitive to the accuracy of measurement instruments.

Continuous probability distributions can be described in several ways. The probability density function describes the infinitesimal probability of any given value, and the probability that the outcome lies in a given interval can be computed by integrating the probabilit density function over that interval. On the other hand, the cumulative distribution function describes the probability that the random variable is no larger than a given value; the probability that the outcome lies in a given interval can be computed by taking the differenc between the values of the cumulative distribution function at the endpoints of the interval. The cumulative distribution function is the antiderivative of the probability density function provided that the latter function exists.

Cumulative distribution function

Because a probability distribution P on the real line is determined by the probability of a scalar random variable X being in a half-open interval $(-\infty, x]$, the probability distribution is completely characterized by its cumulative distribution function:

$$F(x) = \mathrm{P}[X \leq x] \qquad \text{for all } x \in \mathbb{R}.$$

Scanned with CamScanner

Discrete probability distribution

A **discrete probability distribution** is a probability distribution characterized by a probability mass function. Thus, the distribution of a random variable X is discrete, and X is called a **discrete random variable**, if

$$\sum_u \mathrm{P}(X = u) = 1$$

as u runs through the set of all possible values of X. A discrete random variable can assume only a finite or countably infinite number of values. For the number of potential values to be countably infinite, even though their probabilities sum to 1, the probabilities have to decline to zero fast enough. For example, if $\mathrm{P}(X = n) = \frac{1}{2^n}$ for $n = 1, 2, \ldots$, we have the sum of probabilities 1/2 + 1/4 + 1/8 + ... = 1.

Well-known discrete probability distributions used in statistical modeling include the Poisson distribution, the Bernoulli distribution, the binomial distribution, the geometric distribution, and the negative binomial distribution. Additionally, the discrete uniform distribution is commonly used in computer programs that make equal-probability random selections between a number of choices.

Measure theoretic formulation

A measurable function $X: A \to B$ between a probability space $(A, \mathcal{A}, P)$ and a measurable space $(B, \mathcal{B})$ is called a **discrete random variable** provided its image is a countable set and the pre-image of singleton sets are measurable, i.e., $X^{-1}(b) \in \mathcal{A}$ for all $b \in B$. The latter requirement induces a probability mass function $f_X: X(A) \to \mathbb{R}$ via $f_X(b) := P(X^{-1}(b))$. Since the pre-images of disjoint sets are disjoint

$$\sum_{b \in X(A)} f_X(b) = \sum_{b \in X(A)} P(X^{-1}(b)) = P\left(\bigcup_{b \in X(A)} X^{-1}(b)\right) = P(A) = 1.$$

This recovers the definition given above.

Cumulative distribution function

Equivalently to the above, a discrete random variable can be defined as a random variable whose cumulative distribution function (cdf) increases only by jump discontinuities—that is, its cdf increases only where it "jumps" to a higher value, and is constant between those jumps. The points where jumps occur are precisely the values which the random variable may take.

Delta-function representation

Consequently, a discrete probability distribution is often represented as a generalized probability density function involving Dirac delta functions, which substantially unifies the treatment of continuous and discrete distributions. This is especially useful when dealing with probability distributions involving both a continuous and a discrete part.

Indicator-function representation

For a discrete random variable X, let $u_0, u_1, \ldots$ be the values it can take with non-zero probability. Denote

$$\Omega_i = X^{-1}(u_i) = \{\omega : X(\omega) = u_i\}, i = 0, 1, 2, \ldots$$

These are disjoint sets, and for such sets

$$P\left(\bigcup_i \Omega_i\right) = \sum_i P(\Omega_i) = \sum_i P(X = u_i) = 1.$$

It follows that the probability that X takes any value except for $u_0, u_1, \ldots$ is zero, and thus one can write X as

$$X(\omega) = \sum_i u_i 1_{\Omega_i}(\omega)$$

except on a set of probability zero, where 1_A is the indicator function of A. This may serve as an alternative definition of discrete random variables.

Scanned with CamScanner

Continuous probability distribution

A **continuous probability distribution** is a probability distribution that has a cumulative distribution function that is continuous. Most often they are generated by having a probability density function. Mathematicians call distributions with probability density functions **absolutely continuous**, since their cumulative distribution function is absolutely continuous with respect to the Lebesgue measure λ. If the distribution of X is continuous, then X is called a **continuous random variable**. There are many examples of continuous probability distributions: normal, uniform, chi-squared, and others.

Intuitively, a continuous random variable is the one which can take a continuous range of values—as opposed to a discrete distribution, where the set of possible values for the random variable is at most countable. While for a discrete distribution an event with probability zero is impossible (e.g., rolling $3\frac{1}{2}$ on a standard dice is impossible, and has probability zero), this is not so in the case of a continuous random variable. For example, if one measures the width of an oak leaf, the result of 3½ cm is possible; however, it has probability zero because uncountably many other potential values exist even between 3 cm and 4 cm. Each of these individual outcomes has probability zero, yet the probability that the outcome will fall into the interval (3 cm, 4 cm) is nonzero. This apparent paradox is resolved by the fact that the probability that X attains some value within an infinite set, such as an interval, cannot be found by naively adding the probabilities for individual values. Formally, each value has an infinitesimally small probability, which statistically is equivalent to zero.

Formally, if X is a continuous random variable, then it has a probability density function $f(x)$, and therefore its probability of falling into a given interval, say $[a, b]$ is given by the integral

$$\mathrm{P}[a \le X \le b] = \int_a^b f(x)\,dx$$

In particular, the probability for X to take any single value a (that is $a \le X \le a$) is zero, because an integral with coinciding upper and lower limits is always equal to zero.

The definition states that a continuous probability distribution must possess a density, or equivalently, its cumulative distribution function be absolutely continuous. This requirement is stronger than simple continuity of the cumulative distribution function, and there is a special class of distributions, singular distributions, which are neither continuous nor discrete nor a mixture of those. An example is given by the Cantor distribution. Such singular distributions however are never encountered in practice.

Note on terminology: some authors use the term "continuous distribution" to denote the distribution with continuous cumulative distribution function. Thus, their definition includes both the (absolutely) continuous and singular distributions.

By one convention, a probability distribution μ is called *continuous* if its cumulative distribution function $F(x) = \mu(-\infty, x]$ is continuous and, therefore, the probability measure of singletons $\mu\{x\} = 0$ for all x.

Another convention reserves the term *continuous probability distribution* for absolutely continuous distributions. These distributions can be characterized by a probability density function: a non-negative Lebesgue integrable function f defined on the real numbers such that

$$F(x) = \mu(-\infty, x] = \int_{-\infty}^{x} f(t)\,dt.$$

Discrete distributions and some continuous distributions (like the Cantor distribution) do not admit such a density.

Applications

The concept of the probability distribution and the random variables which they describe underlies the mathematical discipline of probability theory, and the science of statistics. There is spread or variability in almost any value that can be measured in a population (e.g. height of people, durability of a metal, sales growth, traffic flow, etc.); almost all measurements are made with some intrinsic error; in physics many processes are described probabilistically, from the kinetic properties of gases to the quantum mechanicaldescription of fundamental particles. For these and many other reasons, simple numbers are often inadequate for describing a quantity, while probability distributions are often more appropriate.

As a more specific example of an application, the cache language models and other statistical language models used in natural language processing to assign probabilities to the occurrence of particular words and word sequences do so by means of probability distributions.

Scanned with CamScanner

Fitting a model

1. How to fit?

1. Both laws and models appear as patterns in data. However it is much easier to spot a law with these patterns than it is to spot a model.
    1. Why should this be? For a law, each data point falls *on* the function that describes the law.
    2. But for a model, each data point falls *around* the function that describes the model. In fact, the data points are *distributed* around the function (as you know).
        1. With enough points, you may be able to notice the structure of these distributions, and therefore make a good guess about the form of the model, but you will rarely have this many points.
        2. More commonly you will need to make the best guess that you can about the model given the data that you have.
        3. In practice, you use an algorithm to select the best guess and R takes care of the details
            1. lm()
            2. xyplot(), makeFun(), plotFun(add = TRUE)
        4. But what is R doing?
2. Many algorithms exist to help you spot the "best" model given the data. In this chapter we will look at how the most popular algorithms work and learn to use them. In the following chapter, we will consider how to choose between the algorithms when modeling a particular data set.

2. Likelihood, the inverse of probability

1. The most intuitive modeling algorithms rely on likelihood. In short, they pick the model that is most *likely* to have generated the data.
2. We use the term likely in everyday speech, but in science *likelihood* has a specific meaning that is closely related to probability.
    1. Probability describes the chance that a certain observation will come to pass given a model. It reasons from model to (future) observation.

SAITEJA XROX

Scanned with CamScanner

2. Likelihood describes the chance that a given model caused the observation that has come to pass. It reasons from observation to model. For this reason, likelihood is sometimes called inverse probability.
    1. Although likelihood and probability both deal with chances they behave differently mathematically. Most notably, the probability of disjoint events must sum to one. Not so for likelihoods.

3. You can calculate the likelihood that a specific model produced a set of data.
    1. Example of a very simple model
        1. goal(y ~ 1, data = ...)
        2. lm()
    2. To find the most likely model compare the likelihoods of different models and choose the model with the highest likelihood.
    3. There is no guarantee that the model with the highest likelihood generated the data, but it is the most pragmatic conclusion to draw. This method of reasoning is known as abduction.

4. In practice, there is a logistical issue to address. There is an infinite number of models to choose from. How will you ever stop calculating likelihoods?
    1. Well, the same way you manage possibilities whenever you make a decision in real life. You first narrow the possibilities down to a smaller set for further consideration. Each modeling algorithm only applies to a small set of models, the set of models that have the structure the algorithm optimizes likelihoods over.
        1. By choosing to use an algorithm, you narrow the set of models to consider to a tractable number. The algorithm then identifies the most likely model within that set.
        2. You should choose the algorithm that matches the structure of the natural law that you wish to model (if you know this structure).
        3. In the next chapter, we will consider how to proceed when you do not know anything about this structure.

SAITEJA XEROX

Scanned with CamScanner

3. Fitting a linear model

1. Linear models are among the oldest and most interpretable modeling methods. A linear model uses a linear function to map a set of values to a set of normal distributions.
    1. Linear models are widely useful because
        1. The normal distribution occurs frequently in the natural world.
        2. Any continuous function can be approximated well with a straight line over a short distance.
    2. Linear models also have another useful feature. For linear models (and generalized linear models with exponential distributions), the model with maximum likelihood will be the model that has the smallest residual sum of squares.
        1. A residual is the distance between the mean of the distribution predicted by a model and the actual data point observed.
        2. Notice that residuals can be positive or negative. Squaring the residuals is a way to measure the magnitude of a residual.
        3. Hence the least squares model will be the model that has the smallest squared residuals, i.e. the model that comes closest to the data points.
2. The lm() function fits a linear model to data.
    1. goal(y ~ x, data = ...)
    2. lm()
3. R's modeling functions return an object that contains a lot of data. To access the data, *store then explore.*
    1. resid()
    2. coef()
    3. fitted()
    4. xyplot(), makeFun(), plotFun(add = TRUE)
4. Linear models are particularly easy to interpret. The coefficient of X is the number of units that the best guess of Y ($\hat{Y}$) increases as X increases by one unit.

G.PRITHVI M.Com, M.Sc, M.B.A, P.G.D.H.M, P.G.D.G.C, P.G.D.C.A

SAI TEJA XEROX

Scanned with CamScanner

4. Discrete terms

1. You can also apply linear models to discrete terms.
    1. Consider this example. To explore the data:
        1. diffmean()
        2. tally()
        3. prop()
        4. perc()
    2. Build your model as you normally would
        1. goal(y ~ 1 | z, data = ...)
        2. lm()
2. Interpretation
    1. R will provide a coefficient for each level of the discrete variable except one. This level will be used as a baseline.
    2. The intercept is the best guess of Y for the baseline group
    3. Each β coefficient is a modifier. It shows how to modify the baseline coefficient to determine the best guess of Y for each remaining group. In other words, it shows the change in the best guess of Y that results from switching from the baseline group to the coefficient's group.
    4. Use factor() to change the baseline group. Your coefficients will change, but the final results will not.

---

5. Multivariate models

1. Add additional terms to the formula to include additional predictors in your model
    1. goal(y ~ x | z, data = ...)
    2. goal(y ~ x + z, data = ...)
2. Now each coefficient should be interpreted as the change in the best guess of Y that results from changing X_i by one unit *while holding the values of the other* *X_j constant*.

SAITEJA XEROX

Scanned with CamScanner

6. Interaction terms

1. Multivariate models create the possibility of interaction effects. An interaction effect occurs when the values of one variable modify the effect of another variable.
    1. A visual explanation
2. Notation
    1. goal(y ~ x + z + x:z, data = ...)
    2. goal(y ~ x*z, data = ...)
3. Interpretation

7. Generalized models

1. lm() uses likelihood to find the model that maps values of *X* to distributions in *Y* that have a *normal*distribution. We saw in Chapter 2 that many natural events will have a normal distribution, but other will not. What if we want to use a model that maps values to non-normal distributions? (This would be appropriate if we are modeling a non-normal *Y* or even a discrete *Y*)
    1. This is an important change because the distributions associated with a model determine how well it fits. They determine how likely the model is to generate the data.
2. It is easy to generalize linear models to non-normal cases by modeling a function of *Y*.
    1. Compare linear model equation to glm model equation
    2. Such functions are known as *link functions*. They map non-normal input (*Y*) to normal output, which can be fitted in the usual way.
    3. These models are known as *generalized linear models (GLM)*
    4. The most common form of generalized linear models are logisitic models
    5. glm(..., family = ...)
        1. options(na.action = "na.exclude")
        2. predict(mod, type = "link"); predict(mod, type = "link")
        3. resid(gmod, type = "deviance"); resid(gmod, type = "pearson")

SAITEJA XEROX

Scanned with CamScanner

3. To interpret a generalized linear model, back transform the coefficients through the link function, or simply try to understand the predictions
    1. Interpreting a logistic regression

---

8. Non-linear models

1. We can use a similar strategy to model non-linear relationships. Instead of modeling Y on X, we can model Y on a function of X.
    1. If you have a particular function of X in mind, you can put it straight into the model equation.
        1. lm()
        2. Don't forget to back transform your results when interpreting
2. *General additive models (GAMS)* fit a model that maps smooth functions of the X_i to distributions in Y. These functions do not need to be linear, only smooth, which makes GAM algorithms useful for mapping many types of relationships.
    1. GAM model equation
    2. library(mgcv)
    3. gam()
    4. s()
        1. interactions
3. As before, we are using likelihood to identify the best model, but our low level tactics have changed. The model with the highest likelihood will be the model whose X functions put the model line exactly through each data point, something that may now be feasible. However, such a model is as unlikely to be true in the practical sense as it is likely to be true in the mathematical sense. As a result, gam() uses a penalized iterative method to select the most likely sensible model.
4. You can combine generalized linear methods and generalized additive methods with gam().
    1. Model equation
    2. Hence, you can think of gam() as a type of generalized modeling algorithm

SAITEJA XEROX

Scanned with CamScanner

5. Interpreting the results of GAMs is difficult. For GAMs as well as other modeling methods that we will encounter later, interpret the model through its predictions.

Intro to R language

R is a programming language and software environment for statistical analysis, graphics representation and reporting. R was created by Ross Ihaka and Robert Gentleman at the University of Auckland, New Zealand, and is currently developed by the R Development Core Team.

R is freely available under the GNU General Public License, and pre-compiled binary versions are provided for various operating systems like Linux, Windows and Mac.

This programming language was named **R**, based on the first letter of first name of the two R authors (Robert Gentleman and Ross Ihaka), and partly a play on the name of the Bell Labs Language S.

Using External Data

R offers plenty of options for loading external data, including Excel, Minitab and SPSS files.

R Session

After R is started, there is a console awaiting for input. At the prompt (>), you can enter numbers and perform calculations.

```
> 1 + 2
[1] 3
```

Variable Assignment

We assign values to variables with the assignment operator "=". Just typing the variable by itself at the prompt will print out the value. We should note that another form of assignment operator "<-" is also in use.

```
> x = 1
> x
[1] 1
```

Functions

R functions are invoked by its name, then followed by the parenthesis, and zero or more arguments. The following apply the function c to combine three numeric values into a vector.

```
> c(1, 2, 3)
[1] 1 2 3
```


SAITEJA XEROX

Scanned with CamScanner

Comments

All text after the pound sign "#" within the same line is considered a comment.

```
> 1 + 1     # this is a comment
[1] 2
```

Extension Package

Sometimes we need additional functionality beyond those offered by the core R library. In order to install an extension package, you should invoke the install.packages function at the prompt and follow the instruction.

```
> install.packages()
```

Getting Help

R provides extensive documentation. For example, entering ?c or help(c) at the prompt gives documentation of the function c in R. Please give it a try.

```
> help(c)
```

If you are not sure about the name of the function you are looking for, you can perform a fuzzy search with the apropos function.

```
> apropos("nova")
[1] "anova"          "anova.glm"
  ....
```

Launching R Programs

In R you can enter each line of code at the prompt in a step-by-step approach. You may also save R programs as simple text files to open in a separate window so that you can enter multiple lines of code at once and save your commands.

Here is an example data set you may save on your computer:

example1.dat

Here is an example program:

```
#Read data file into R as a vecto
example1 = scan("/Users/Shared/WD/Rdirectory/example1.dat")
#Change pathname to wherever you saved example1.d
```


SAITEJA XEROX

Scanned with CamScanner

```
#Print data
> example1
#Calculate the sample mean
> mean(example1)
```

Data

Depending on the course, datasets are either presented within the context of the lesson or within a datasets folder. Common file extensions for data files include ".dat", ".csv", and ".txt". You must download the data from your course website. Canvas provides instructions on how to save a file for Windows users or Mac users. A Save dialog box will be displayed and allow you to save the data file to the location you choose on your computer.

There are a number of ways to read data into your R session. Two popular commands used in the examples presented here are read.table and scan.

UNIT – II

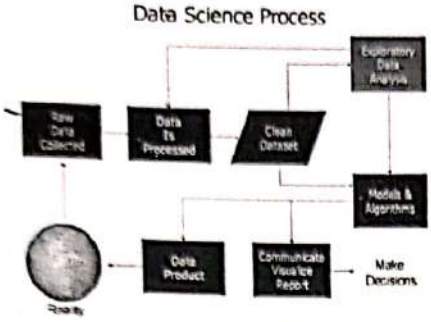
Exploratory Data analysis and the Data science process

Exploratory data analysis (EDA) is an approach to analyzing data sets to summarize their main characteristics, often with visual methods. A statistical model can be used or not, but primarily EDA is for seeing what the data can tell us beyond the formal modeling or hypothesis testing task. Exploratory data analysis was promoted by John Tukey to encourage statisticians to explore the data, and possibly formulate hypotheses that could lead to new data collection and experiments. EDA is different from initial data analysis (IDA), which focuses more narrowly on checking assumptions required for model fitting and hypothesis testing, and handling missing values and making transformations of variables as needed. EDA encompasses IDA.

Tukey defined data analysis in 1961 as: "[P]rocedures for analyzing data, techniques for interpreting the results of such procedures, ways of planning the gathering of data to make its analysis easier, more precise or more accurate, and all the machinery and results of (mathematical) statistics which apply to analyzing data."

SAITHJA XEROX

Scanned with CamScanner



Basic Tools of EDA

(Plots, graphs and summary statistics)

Plot

A **plot** is a graphical technique for representing a data set, usually as a graph showing the relationship between two or more variables. The plot can be drawn by hand or by a mechanical or electronic plotter. Graphs are a visual representation of the relationship between variables, very useful for humans who can quickly derive an understanding which would not come from lists of values. Graphs can also be used to read off the value of an unknown variable plotted as a function of a known one. Graphs of functions are used in mathematics, sciences, engineering, technology, finance, and other areas.

Plots play an important role in statistics and data analysis. The procedures here can broadly be split into two parts: quantitative and graphical. Quantitative techniques are the set of statistical procedures that yield numeric or tabular output. Examples of quantitative techniques include:

- hypothesis testing
- analysis of variance
- point estimates and confidence intervals
- least squares regression

These and similar techniques are all valuable and are mainstream in terms of classical analysis. There are also many statistical tools generally referred to as graphical techniques. These include:[1]

- scatter plots
- histograms
- probability plots
- residual plots

SATTEJAXEROX

Scanned with CamScanner

- box plots, and
- block plots

Graphical procedures such as plots are a short path to gaining insight into a data set in terms of testing assumptions, model selection, model validation, estimator selection, relationship identification, factor effect determination, outlier detection. Statistical graphics give insight into aspects of the underlying structure of the data.

Graphs can also be used to solve some mathematical equations, typically by finding where two plots intersect.

Graph

A **graph** is a structure amounting to a set of objects in which some pairs of the objects are in some sense "related". The objects correspond to mathematical abstractions called *vertices* (also called *nodes* or *points*) and each of the related pairs of vertices is called an *edge* (also called an *arc* or *line*). Typically, a graph is depicted in diagrammatic form as a set of dots for the vertices, joined by lines or curves for the edges. Graphs are one of the objects of study in discrete mathematics.

The edges may be directed or undirected. For example, if the vertices represent people at a party, and there is an edge between two people if they shake hands, then this graph is undirected because any person A can shake hands with a person B only if B also shakes hands with A. In contrast, if any edge from a person A to a person B corresponds to A's admiring B, then this graph is directed, because admiration is not necessarily reciprocated. The former type of graph is called an *undirected graph* and the edges are called *undirected edges* while the latter type of graph is called a *directed graph* and the edges are called *directed edges*.

In one very common sense of the term, a *graph* is an ordered pair $G = (V, E)$ comprising a set V of *vertices, nodes* or *points* together with a set E of *edges, arcs* or *lines*, which are 2-element subsets of V (i.e., an edge is associated with two vertices, and the association takes the form of the unordered pair of the vertices). To avoid ambiguity, this type of graph may be described precisely as *undirected* and *simple*.

Other senses of *graph* stem from different conceptions of the edge set. In one more general conception, E is a set together with a relation of *incidence* that associates with each edge two vertices. In another generalized notion, E is a multiset of unordered pairs of (not necessarily distinct) vertices. Many authors call these types of object multigraphs or pseudographs.

All of these variants and others are described more fully below.

SAI TEJA XEROX

Scanned with CamScanner

The vertices belonging to an edge are called the *ends* or *end vertices* of the edge. A vertex may exist in a graph and not belong to an edge.

V and E are usually taken to be finite, and many of the well-known results are not true (or are rather different) for *infinite graphs* because many of the arguments fail in the infinite case. Moreover, V is often assumed to be non-empty, but E is allowed to be the empty set. The *order* of a graph is $|V|$, its number of vertices. The *size* of a graph is $|E|$, its number of edges. The *degree* or *valency* of a vertex is the number of edges that connect to it, where an edge that connects to the vertex at both ends (a loop) is counted twice.

For an edge $\{x, y\}$, graph theorists usually use the somewhat shorter notation xy.

Adjacency relation

The edges E of an undirected graph G induce a symmetric binary relation $\sim$ on V that is called the *adjacency relation* of G. Specifically, for each edge $\{x, y\}$, the vertices x and y are said to be *adjacent* to one another, which is denoted $x \sim y$.

Summary statistics

summary statistics are used to summarize a set of observations, in order to communicate the largest amount of information as simply as possible. Statisticians commonly try to describe the observations in

- a measure of location, or central tendency, such as the arithmetic mean
- a measure of statistical dispersion like the standard deviation
- a measure of the shape of the distribution like skewness or kurtosis
- if more than one variable is measured, a measure of statistical dependence such as a correlation coefficient

A common collection of order statistics used as summary statistics are the five-number summary, sometimes extended to a seven-number summary, and the associated box plot.

Entries in an analysis of variance table can also be regarded as summary statistics.

Examples of summary statistics

Location

Common measures of location, or central tendency, are the arithmetic mean, median, mode, and interquartile mean.

SAI TEJA XEROX
Scanned with CamScanner

Spread

Common measures of statistical dispersion are the standard deviation, variance, range, interquartile range, absolute deviation, mean absolute difference and the distance standard deviation. Measures that assess spread in comparison to the typical size of data values include the coefficient of variation.

The Gini coefficient was originally developed to measure income inequality and is equivalent to one of the L-moments.

A simple summary of a dataset is sometimes given by quoting particular order statistics as approximations to selected percentiles of a distribution.

Shape

Common measures of the shape of a distribution are skewness or kurtosis, while alternatives can be based on L-moments. A different measure is the distance skewness, for which a value of zero implies central symmetry.

Dependence

The common measure of dependence between paired random variables is the Pearson product-moment correlation coefficient, while a common alternative summary statistic is Spearman's rank correlation coefficient. A value of zero for the distance correlation implies independence.

Philosophy of EDA

Exploratory data analysis (EDA) was introduced by John Tukey as an approach to analyze data when there is only a low level of knowledge about its cause system as well as *contextual*information. EDA aims at letting the data itself influence the process of suggesting hypotheses instead of only using it to evaluate given *(a priori)* hypotheses.

Explorative - opposed to Confirmatory - Data Analysis is like detective work looking for patterns, anomalies or in general new insights and is usually done via graphical representations of the underlying data-set.

Exploratory Data Analysis (EDA) is detective work – numerical detective work – or counting detective work – or graphical detective work ... unless exploratory data analysis uncovers indications, usually quantitative ones, there is likely to be nothing for confirmatory data analysis to consider ... [it] can never be the whole story, but nothing else can serve as the foundation stone - as the first step.

[Tukey, 1977]

SAITEJA XEROX

Scanned with CamScanner

- *Exploratory Data Analysis (EDA) is an approach/philosophy for data analysis that employs a variety of techniques (mostly graphical) to maximize*

1. insight into a data set;

2. uncover underlying structure;

3. extract important variables;

4. detect outliers and anomalies;

5. test underlying assumptions;

6. develop parsimonious models; and

7. determine optimal factor settings.

The EDA approach is precisely that--an approach--not a set of techniques, but an attitude/philosophy about how a data analysis should be carried out.

[Filliben, 2004]

[...] is concerned primarily with explorations and description of data, not with inference. The techniques are designed to identify fundamental, conceptually meaningful patterns and relationships in data and to call attention to observations that deviate greatly from those fundamental patterns

[Smith and Prentice, 1993]

Exploratory data analysis as the process of searching and analyzing databases to find implicit but potentially useful information, is a difficult task. At the beginning, the analyst has no hypothesis about the data. According to John Tuckey, tools as well as understanding are needed [Tukey, 1977] for the interactive and usually undirected search for structures and trends.

[Keim et al., 2006]

Furthermore, EDA can be used to support the selection of appropriate statistical tools as well as to provide a basis for statistical inference and further data collection.

Essential to EDA are graphical tools like *box plots, stem–and–leaf plots, scatter plots*, or timelines.

John W. Tukey wrote the book *Exploratory Data Analysis* in 1977. Tukey held that too much emphasis in statistics was placed on statistical hypothesis testing (confirmatory data analysis); more emphasis needed to be placed on using datatosuggest hypotheses to test. In particular, he held that confusing the two types of analyses and employing them on the same set of data can lead to systematic bias owing to the issues inherent in testing hypotheses suggested by the data.

SAITEJA XEROX

Scanned with CamScanner

The objectives of EDA are to:

- Suggest hypotheses about the causes of observed phenomena
- Assess assumptions on which statistical inference will be based
- Support the selection of appropriate statistical tools and techniques
- Provide a basis for further data collection through surveys or experiments

Many EDA techniques have been adopted into data mining, as well as into big data analytics. They are also being taught to young students as a way to introduce them to statistical thinking.

Techniques

There are a number of tools that are useful for EDA, but EDA is characterized more by the attitude taken than by particular techniques.

Typical graphical techniques used in EDA are:

- Box plot
- Histogram
- Multi-vari chart
- Run chart
- Pareto chart
- Scatter plot
- Stem-and-leaf plot
- Parallel coordinates
- Odds ratio
- Targeted projection pursuit
- Dimensionality reduction:
    - Multidimensional scaling
    - Principal component analysis (PCA)
    - Multilinear PCA
    - Nonlinear dimensionality reduction (NLDR)

SAI TEJA XEROX

Scanned with CamScanner

- Projection methods such as grand tour, guided tour and manual tour
- Interactive versions of these plots

Typical quantitative techniques are:

- Median polish
- Trimean
- Ordination

The Data science process

Step 1: Frame the problem

The first thing you have to do before you solve a problem is to define exactly what it is. You need to be able to translate data questions into something actionable.

You'll often get ambiguous inputs from the people who have problems. You'll have to develop the intuition to turn scarce inputs into actionable outputs–and to ask the questions that nobody else is asking.

Say you're solving a problem for the VP Sales of your company. You should start by understanding their goals and the underlying *why* behind their data questions. Before you can start thinking of solutions, you'll want to work with them to clearly define the problem.

A great way to do this is to ask the right questions.

You should then figure out what the sales process looks like, and who the customers are. You need as much context as possible for your numbers to become insights.

You should ask questions like the following:

1. Who are the customers?
2. Why are they buying our product?
3. How do we predict if a customer is going to buy our product?
4. What is different from segments who are performing well and those that are performing below expectations?
5. How much money will we lose if we don't actively sell the product to these groups?

In response to your questions, the VP Sales might reveal that they want to understand why certain segments of customers have bought less than expected. Their end goal might be to determine whether to

SAITEJA XEROX

Scanned with CamScanner

continue to invest in these segments, or de-prioritize them. You'll want to tailor your analysis to that problem, and unearth insights that can support either conclusion.

It's important that at the end of this stage, you have all of the information and context you need to solve this problem.

Step 2: Collect the raw data needed for your problem

Once you've defined the problem, you'll need data to give you the insights needed to turn the problem around with a solution. This part of the process involves thinking through what data you'll need and finding ways to get that data, whether it's querying internal databases, or purchasing external datasets.

You might find out that your company stores all of their sales data in a CRM or a customer relationship management software platform. You can export the CRM data in a CSV file for further analysis.

Step 3: Process the data for analysis

Now that you have all of the raw data, you'll need to process it before you can do any analysis. Oftentimes, data can be quite messy, especially if it hasn't been well-maintained. You'll see errors that will corrupt your analysis: values set to null though they really are zero, duplicate values, and missing values. It's up to you to go through and check your data to make sure you'll get accurate insights.

You'll want to check for the following common errors:

1. Missing values, perhaps customers without an initial contact date
2. Corrupted values, such as invalid entries
3. Timezone differences, perhaps your database doesn't take into account the different timezones of your users
4. Date range errors, perhaps you'll have dates that makes no sense, such as data registered from before sales started

You'll need to look through aggregates of your file rows and columns and sample some test values to see if your values make sense. If you detect something that doesn't make sense, you'll need to remove that data or replace it with a default value. You'll need to use your intuition here: if a customer doesn't have an initial contact date, does it make sense to say that there was NO initial contact date? Or do you have to hunt down the VP Sales and ask if anybody has data on the customer's missing initial contact dates?

Once you're done working with those questions and cleaning your data, you'll be ready for exploratory data analysis (EDA).

G.PRITHVI M.Com, M.Sc, M.B.A, P.G.D.H.M, P.G.D.G.C, P.G.D.C.A

SAI TEJA XEROX

Scanned with CamScanner

Step 4: Explore the data

When your data is clean, you'll should start playing with it!

The difficulty here isn't coming up with ideas to test, it's coming up with ideas that are likely to turn into insights. You'll have a fixed deadline for your data science project (your VP Sales is probably waiting on your analysis eagerly!), so you'll have to prioritize your questions.

You'll have to look at some of the most interesting patterns that can help explain why sales are reduced for this group. You might notice that they don't tend to be very active on social media, with few of them having Twitter or Facebook accounts. You might also notice that most of them are older than your general audience. From that you can begin to trace patterns you can analyze more deeply.

Step 5: Perform in-depth analysis

This step of the process is where you're going to have to apply your statistical, mathematical and technological knowledge and leverage all of the data science tools at your disposal to crunch the data and find every insight you can.

In this case, you might have to create a predictive model that compares your underperforming group with your average customer. You might find out that the age and social media activity are significant factors in predicting who will buy the product.

If you'd asked a lot of the right questions while framing your problem, you might realize that the company has been concentrating heavily on social media marketing efforts, with messaging that is aimed at younger audiences. You would know that certain demographics prefer being reached by telephone rather than by social media. You begin to see how the way the product has been has been marketed is significantly affecting sales: maybe this problem group isn't a lost cause! A change in tactics from social media marketing to more in-person interactions could change everything for the better. This is something you'll have to flag to your VP Sales.

You can now combine all of those qualitative insights with data from your quantitative analysis to craft a story that moves people to action.

Step 6: Communicate results of the analysis

It's important that the VP Sales understand why the **insights you've uncovered are important**. Ultimately, you've been called upon to create a solution throughout the data science process. **Proper communication** will mean the difference between **action** and **inaction** on your proposals.

You need to craft a compelling story here that ties your data with their knowledge. You start by explaining the reasons behind the underperformance of the older demographic. You tie that in with the

SAITEJA XEROX

Scanned with CamScanner

answers your VP Sales gave you and the insights you've uncovered from the data. Then you move to concrete solutions that address the problem: we could shift some resources from social media to personal calls. You tie it all together into a narrative that solves the pain of your VP Sales: she now has clarity on how she can reclaim sales and hit her objectives.

She is now ready to act on your proposals.

Throughout the data science process, your day-to-day will vary significantly depending on where you are—and you will definitely receive tasks that fall outside of this standard process! You'll also often be juggling different projects all at once.

It's important to understand these steps if you want to systematically think about data science, and even more so if you're looking to start a career in data science.

Case study: RealDirect (Online Real estate firm)

Doug Perlson, the CEO of RealDirect, has a background in real estate law, startups, and online advertising. His goal with RealDirect is to use all the data he can access about real estate to improve the way people sell and buy houses. Normally, people sell their homes about once every seven years, and they do so with the help of professional brokers and current data. But there's a problem both with the broker system and the data quality. RealDirect addresses both of them.

First, the brokers. They are typically "free agents" operating on their own – think of them as home sales consultants. This means that they guard their data aggressively, and the really good ones have lots of experience. But in the grand scheme of things, that really means they have only slightly more data than the inexperienced brokers. RealDirect is addressing this problem by hiring a team of licensed real estate agents who work together and pool their knowledge. To accomplish this, it built an interface for sellers, giving them useful data driven tips on how to sell their house. It also uses interaction data to give real-time recommendations on what to do next.

The team of brokers also become data experts, learning to use information-collecting tools to keep tabs on new and relevant data or to access publicly available information. For example, you can now get data on co-op (a certain kind of apartment in NYC) sales, but that's a relatively recent change.

One problem with publicly available data is that it's old news—there's a three-month lag between a sale and when the data about that sale is available. RealDirect is working on real-time feeds on things like when people start searching for a home, what the initial offer is, the time between offer and close, and how people search for a home online. Ultimately, good information helps both the buyer and the seller. At least if they're honest.

SAITEJA XEROX

Scanned with CamScanner

How Does RealDirect Make Money?

First, it offers a subscription to sellers—about $395 a month—to access the selling tools. Second, it allows sellers to use RealDirect's agents at a reduced commission, typically 2% of the sale instead of the usual 2.5% or 3%. This is where the magic of data pooling comes in: it allows RealDirect to take a smaller commission because it's more optimized, and therefore gets more volume.

The site itself is best thought of as a platform for buyers and sellers to manage their sale or purchase process. There are statuses for each person on site: active, offer made, offer rejected, showing, in contract, etc. Based on your status, different actions are suggested by the software.

There are some challenges they have to deal with as well, of course. First off, there's a law in New York that says you can't show all the current housing listings unless those listings reside behind a registration wall, so RealDirect requires registration. On the one hand, this is an obstacle for buyers, but serious buyers are likely willing to do it. Moreover, places that don't require registration, like Zillow, aren't true competitors to RealDirect because they are merely showing listings without providing any additional service. Doug pointed out that you also need to register to use Pinterest, and it has tons of users in spite of this.

RealDirect comprises licensed brokers in various established realtor associations, but even so it has had its share of hate mail from realtors who don't appreciate its approach to cutting commission costs. In this sense, RealDirect is breaking directly into a guild. On the other hand, if a realtor refused to show houses because they are being sold on RealDirect, the potential buyers would see those listings elsewhere and complain. So the traditional brokers have little choice but to deal with RealDirect even if they don't like it. In other words, the listings themselves are sufficiently transparent so that the traditional brokers can't get away with keeping their buyers away from these houses.

Doug talked about key issues that a buyer might care about—nearby parks, subway, and schools, as well as the comparison of prices per square foot of apartments sold in the same building or block. This is the kind of data they want to increasingly cover as part of the service of RealDirect.

Exercise: RealDirect Data Strategy

You have been hired as a Business Analytics Consultant at *realdirect.com*, and report directly to the CEO. The company (hypothetically) does not yet have its data plan in place. It's looking to you to come up with a data strategy.

Here are a couple ways you could begin to approach this problem:

SAITEJA XEROX

Scanned with CamScanner

1. Explore its existing website, thinking about how buyers and sellers would navigate through it, and how the website is structured/organized. Try to understand the existing business model, and think about how analysis of RealDirect user-behavior data could be used to inform decision-making and product development. Come up with a list of research questions you think could be answered by data:

- What data would you advise the engineers log and what would your ideal datasets look like?
- How would data be used for reporting and monitoring product usage?
- How would data be built back into the product/website?

2. Because there is no data yet for you to analyze (typical in a startup when its still building its product), you should get some auxiliary data to help gain intuition about this market.

- First challenge: load in and clean up the data. Next, conduct exploratory data analysis in order to find out where there are outliers or missing values, decide how you will treat them, make sure the dates are formatted correctly, make sure values you think are numerical are being treated as such, etc.
- Once the data is in good shape, conduct exploratory data analysis to visualize and make comparisons (i) **across neighborhoods**, and (ii) **across time**. If you have time, start looking for meaningful patterns in this dataset.

3. Summarize your findings in a brief report aimed at the CEO.

4. Being the "business analyst" often involves speaking to people who aren't also analysts, so it would be ideal to have a set of communication strategies for getting to the information you need about the data. Can you think of any other people you should talk to?

5. Most of you are not "domain experts" in real estate or online businesses.

- Does stepping out of your comfort zone and figuring out how you would go about "collecting data" in a different setting give you insight into how you do it in your own field?
- Sometimes "domain experts" have their own set of vocabulary. Did Doug use vocabulary specific to his domain that you didn't understand ("comps," "open houses," "CPC")? Sometimes if you don't understand vocabulary that an expert is using, it can prevent you from understanding the problem. It's good to get in the habit of asking questions because eventually you will get to something you do understand. This involves persistence and is a habit to cultivate.

SAITEJA XEROX

Scanned with CamScanner

6. Doug mentioned the company didn't necessarily have a data strategy. There is no industry standard for creating one. As you work through this assignment, think about whether there is a set of best practices you would recommend with respect to developing a data strategy for an online business, or in your own domain.

UNIT – III

Feature Generation and Feature Selection (Extracting Meaning from data)

Feature generation

Feature generation is also known as feature construction, feature extraction or feature engineering. There are different interpretations of the terms feature generation, construction, extraction and engineering. Some nearly equivalent, yet differing definitions for these terms are; construction of features from raw data [11, 6], creating a mapping to convert original features to new features [12], creating new features from one or multiple features [10].

In the context of this thesis, feature generation and all its synonyms will be interpreted as the process of creating new features from one or multiple features, unless otherwise specified.

Two goals of feature generation can be dimensionality reduction and accuracy improvement [13, 4]. When the goal of a feature generation method is dimensionality reduction, then the result will be a feature space which contains less features than the original feature space. However, when the goal is accuracy improvement, the resulting feature space will most likely contain more features than the original feature space.

We are primarily interested in feature generation methods where the goal is to improve the accuracy of the predictor. Dimensionality reduction does not have a high priority, since the results of feature generation are input to a feature selection phase aims to reduce the dimensionality of the feature space. Even though the feature generation phase does not have to reduce the dimensionality, it certainly has to take care to not generate an extreme amount of new features.

To illustrate the importance of feature generation, consider the following example in Table 1.1. Here we can see the original feature `Date` and dependent feature `Visitors`, which shows a date and the corresponding number of visitors for a theme park. When looking at just these features, there does not seem to be an obvious pattern between the predicting and the dependent feature. With feature generation, we can extract what kind of day the date is, shown in the `IsWeekendDay` column. This tells us whether the date is a weekend day or not. Now we can see a clear pattern where the number of visitors is significantly higher on weekend days than on week days.

Scanned with CamScanner

Date	Visitors	IsWeekendDay
'20-05-17'	19234	yes
'11-04-17'	5735	no
'29-01-17'	17914	yes
'04-05-17'	5496	no
'29-03-17'	5913	no

Table 1.1: Example features Day and Visitors and extracted feature IsWeekendDay{Day}

Another situation where feature generation can improve performance is when there is feature interaction. Then two (or more) features are not relevant or correlated to the dependent feature on their own, but together they have a (high) influence on the dependent feature. For example, take as features the price and quality of a product. Separately, they will not give much indication of whether a product is purchased often. Combined they have a high correlation to the purchase of the product. If the price is low and the quality high, then the product will be purchased often. However, a low price or a high quality without knowing the other value cannot guarantee that the product will be purchased often. If both price and quality are low, then the product will not be purchased by many customers. The same can be said when both price and quality are high.

SAI TEJA

Scanned with CamScanner

Feature selection

Feature selection tries to find the optimal subset of a given feature set. The problem of feature selection is essentially equivalent to the problem of finding the optimal subset of a given set, which has been shown to be NP-hard. There is one simple method to find the optimal subset, namely calculate the evaluation score for each possible subset. The advantage of this method is that it will always find the optimal subset. A disadvantage is that the complexity is $O(2^n)$, where n is the number of features. This means that for 10 features, already 1024 sets have to be evaluated. This disadvantage is even more influential, because feature selection will become more useful the more features are involved.

Since evaluating every subset separately is practically infeasible, there is need for a different, smarter way to decide the usefulness of each subset.

If the original feature set contains a thousand features, it is highly likely that not all features positively influence the dependent variable. For example, in text mining it is common practice to remove the stop words; the, in, a, of, as, and, with, and many more. Although stop word removal can be seen as a part of the data cleaning step, it can also be seen as a separate feature selection step. Consider every word to be a feature, then the stop words are features that do not contribute to the prediction of the dependent value.

Feature irrelevance is the problem that some features are simply not correlated to the dependent feature. These features can even have a negative effect on the performance of a model.

Data extraction is where data is analyzed and crawled through to retrieve relevant information from data sources (like a database) in a specific pattern. Further data processing is done, which involves adding metadata and other data integration; another process in the data workflow.

The majority of data extraction comes from unstructured data sources and different data formats. This unstructured data can be in any form, such as tables, indexes, and analytics.

Motivating application: user (customer) retention

For most companies, the cost of acquiring a new customer is far more than the cost of retaining an existing customer—often 5-10 times more expensive. Moreover, 61% of small business owners have reported that more than half of their annual revenue comes from repeat buyers.

Those impressive numbers confirm the 80/20 rule to be true (20% of customers bring 80% of the business); and, thanks to the substantial amount of customer data businesses now have available, many businesses are shifting their primary focus to these customers.

If a business is to retain current customers, then its marketers must truly understand repeat customers' needs to improve their overall experience with the brand and gain their long-term loyalty.

Scanned with CamScanner

For most marketers, it is no longer a challenge to collect customer data: Analytical technologies have given us the tools to understand a customer's actions at every point of interaction with a brand. But many marketers are still struggling to transform that analytical data into relevant information that can help improve customer loyalty.

Fortunately, there are steps marketers can take to harness data and keep churn rates at a minimum.

Companies have been using data science as a secret weapon to generate quickly actionable information that improves customer retention. If you're interested in significantly decreasing your customer churn rate, here's how to use the power of data science to define a process that will help your brand keep the customers you've already worked so hard to obtain.

Step 1: To identify churners, define your business model

Nowadays, subscription and recurring-revenue business models are everywhere. And no wonder: how customers want to access and pay for goods and services is changing, so companies are in turn changing their pricing models.

The first step in setting up a data-driven approach to increasing customer loyalty is to define the model used by your organization. More than likely, your model will fall under one of two varieties: a subscription model or a non-subscription model. (Netflix and Spotify are examples of a subscription model; Uber and eBay are examples of a non-subscription model.)

Your business model has an impact on the difficulty of determining "churn rate," which can have different definitions; however, for most businesses, it's a matter of whether a customer will become a "churner"—i.e., no longer a customer. Churn could also refer to the loss of contracts, MRR (monthly recurring revenue), contract value, and bookings.

Churn is frequently expressed as a rate, a ratio, or a whole number: for example, "We have a churn rate of 10%," and, "We churned five customers."

Identifying churners is straightforward in a subscription model: A customer churns when she requests a cancellation of her subscription. In a non-subscription model, however, you will need to analyze a customer's behavioral tendencies to identify possible churn—such as the amount of time since he or she last used your company's services. The primary goal is to determine the specific point after which your customer will no longer use your product or service.

SAITEJA XEROX

Scanned with CamScanner

Step 2: Decide on an approach to retain potential churners

Now that you've determined your business model, there are many ways you can start retaining potential churners with both short-term actions and a long-term approach. To have the most effective and sustainable impact on customer loyalty, you can try a combination of the following two methods.

1. One-shot, short-term actions to reduce churn

- Launch special offers in the form of calls, push notifications, free in-game money, discount coupons, etc.
- To control or measure customer satisfaction, set up customer feedback loops in the form of surveys, on-website or in-app questionnaires, self-service options such as designated review or feedback sections.

2. Longer-term approaches to attack the root of a problem to reduce churn

- Reduce obstacles to make it easier to purchase your product.
- Analyze whether your the offer fits your customer base.

Step 3: Create customer profiles (segments) and determine their behavior

What do you need to provide the right offer to your customers at the right moment (i.e., when they are making a purchasing decision)?

The answer: in-depth knowledge about that customer. However, too often customers avoid handing over personal information that can be useful to marketers—such as age, gender, profession, and buying habits. That's why a data-driven approach is an advantage.

Instead of asking customers for such information, it can be collected and assessed externally. The success of your business is directly linked to a true understanding of your customers, and a data science approach works only when you have a defined target. In a subscription business model, the target is known; however, if you're operating a non-subscription model, you will first need to define your target if you're going to understand it.

You can create a customer profile and identify behaviors by asking yourself (or your team) the following questions:

- **Which customers do we care about?** Segment your customers based on their behavior, and ask, "Which customers do we care most about?" Regardless of the answer, the only way to increase customer loyalty with this type of campaign is to target a well-defined customer segment.

SAITEJA XEROX

Scanned with CamScanner

- **How will we segment new customers?** By understanding the extremes of your customers, you can create and refine new customer segments. On one extreme, you will have customers who interacted with your brand at least once, but discontinued interaction afterward. The other extreme will include customers who use your product or service frequently or are heavily engaged with your brand.

  Once you identify these extremes and understand your customer segmentations, it will be much easier to place a new customer in the appropriate grouping.

- **What makes our churner customers different?** You will begin identifying patterns among your churners and define what makes them different fomr others. Make note of, and discuss, these differences frequently.

Step 4: Define and implement a customer-scoring process

Now that you've identified your target customers and understand their behaviors, you can implement a method of scoring your customers based on the data you have about them.

Customer traits, such as social information and behavior-based actions, can be used to paint a picture of who they are. Then, you can compute a score that incorporates all relevant customer features to determine exactly how likely that person is to abandon your offerings.

This is the point at which you can harness data science to take over. A predictive analytics platform can be used to aggregate all your customer data, identify potential churners, and calculate a score that predicts the potential loyalty of your customers.

This type of scoring system enables marketers to creatively segment customers and activate the most appropriate marketing strategies for each. For example, potential customer segments and actions for each could be laid out like this:

- Loyal customers: Take no action.
- Potential churners (customers you want to keep): Send a special offer via email.
- Churners: Send a special offer via email.
- Ambivalent (unsure whether to keep): Send a simple greeting without an offer.

Feature Generation

This is the process of taking raw, unstructured data and defining features (i.e. variables) for potential use in your statistical analysis. For instance, in the case of text mining you may begin with a raw log of

SAITEJA XEROX

Scanned with CamScanner

thousands of text messages (e.g. SMS, email, social network messages, etc) and generate features by removing low-value words (i.e. stopwords), using certain size blocks of words (i.e. n-grams) or applying other rules.

Brainstorming

Brainstorming is a group creativity technique by which efforts are made to find a conclusion for a specific problem by gathering a list of ideas spontaneously contributed by its members.

The term was popularized by Alex Faickney Osborn in the 1953 book *Applied Imagination.*

Osborn's method

Osborn claimed that two principles contribute to "ideative efficacy," these being :

1. Defer judgment,
2. Reach for quantity.

Following these two principles were his four general rules of brainstorming, established with intention to :

- reduce social inhibitions among group members.
- stimulate idea generation.
- increase overall creativity of the group.

1. **Go for quantity:** This rule is a means of enhancing divergent production, aiming to facilitate problem solving through the maxim *quantity breeds quality*. The assumption is that the greater the number of ideas generate the bigger the chance of producing a radical and effective solution.
2. **Withhold criticism:** In brainstorming, criticism of ideas generated should be put 'on hold'. Instead, participants should focus on extending or adding to ideas, reserving criticism for a later 'critical stage' of the process. By suspending judgment, participants will feel free to generate unusual ideas.
3. **Welcome wild ideas:** To get a good long list of suggestions, wild ideas are encouraged. They can be generated by looking from new perspectives and suspending assumptions. These new ways of thinking might give you better solutions.
4. **Combine and improve ideas:** As suggested by the slogan "1+1=3". It is believed to stimulate the building of ideas by a process of association.

SAITEJA XEROX

Scanned with CamScanner

Applications

Osborn notes that brainstorming should address a specific question; he held that sessions addressing multiple questions were inefficient.

Further, the problem must require the generation of ideas rather than judgment; he uses examples such as generating possible names for a product as proper brainstorming material, whereas analytical judgments such as whether or not to marry do not have any need for brainstorming.

Groups

Osborn envisioned groups of around 12 participants, including both experts and novices. Participants are encouraged to provide wild and unexpected answers. Ideas receive no criticism or discussion. The group simply provides ideas that might lead to a solution and apply no analytical judgment as to the feasibility. The judgments are reserved for a later date

In the world of business and technology, success stems from good ideas. Successful organizations rely on a variety of idea generation techniques to:

- Solve problems
- Identify options, needs and features
- Improve processes
- Resolve issues
- Create or modify products

Teams assume good idea generation techniques inspire collaboration, creativity and innovation, but struggle to transform those assumptions to reality. They don't find the promised value in their brainstorming sessions. Instead, meetings end with stakeholders buried in piles of sticky notes wondering where the value is hiding.

Stakeholders forget that idea generation is only the first step—maybe the easiest step. Piles of sticky notes and rolled up flip chart sheets don't offer value on their own. We need to find a way to process the output of your brainstorming session until a clear path to value appears.

Process The Output: Divergent And Convergent Thinking

Good idea generation techniques require divergent thinking where all participants submit new ideas and expand existing ideas without criticism, constraints or editing. Words like impossible, impractical, stupid, too expensive do not exist in this divergent process.

SAI TEJA XEROX

Scanned with CamScanner

This process makes some team members uncomfortable. BAs need to let visions of business rules, risks, process diagrams, and system constraints fade away. Concerns about scope and cost do not apply in this open brainstorming phase.

(Some companies use squirt guns to quiet people who critique divergent ideas.)

After the divergent phase, the real work begins. How do you process the ideas? How do you narrow them down to something tangible and valuable? Here are a few possibilities for the first convergent thinking phase:

- Condense-remove duplicates, group similar ideas
- Categorize-Often categories or themes emerge that can be an umbrella for multiple ideas
- Matrix-Ideas may fit into a matrix like SWOT (strengths, weaknesses, opportunities, threats) or high/low cost vs. high/low time, high/low benefit vs. high/low complexity
- Range/Scale-List ideas in order by feasibility, practicality, cost, time, impact or value.

A Clear Path: Prioritization

After the initial convergent thinking phase, an idea prioritization process provides a clear path to value. There are many ways to prioritize a list of ideas and the proper technique will vary depending on the nature of your ideas, but here are three suggestions:

- Voting-Allow participants to vote on their favorite ideas virtually with a survey tool or in-person with tallies or dot stickers. Consider the value of anonymous voting. Consider allowing more than one vote. Consider allowing no votes and yes votes.
- Buy a Feature – a prioritization technique that allows stakeholders to vote and prioritize with "play money". Encourage customers to pool their money to buy important/expensive features.
- Forced Ranking - Place two ideas/features in front of the group, ask the group to decide if the second idea goes above or below the first. (Ideas at the top of the list have the highest priority.) Then take the third idea, ask the group if it goes above, below or between the others. Continue through all ideas until they are ranked in order.

Finding Value: Taking Action

The prioritized list of ideas should offer a clear path for action.

- If the list includes new features for a product, the team might start defining requirements for the top features.

SAITEJA XEROX

Scanned with CamScanner

- If the list includes solutions for a problem, the team might attempt the solutions in order of priority until they found success.
- If the list includes ideas for process improvements, the team might begin process change procedures for the top requirements.

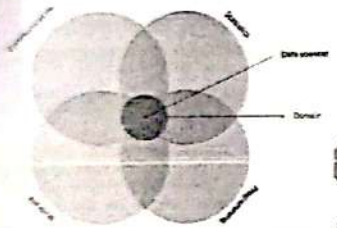
When working through the priorities remember to check-in with stakeholders to reassess. New priorities might be added and existing priorities might shift.

Role of domain expertise

Over the last several years, there has (rightly) been significant debate on the skill sets of data scientists, and the importance of business, statistics, programming and other skill sets. Interesting sub-classifications of professions, such as "data hacker", "data nerd" and other terms have been used to describe the various combinations or intersections of these skill sets.

THE IMPORTANCE OF DOMAIN KNOWLEDGE

In all of these discussions, however, one key element has been left out. And that is the **domain**.



Domain knowledge is an important subset of the data scientist's work. Although the perfect data scientist is a bit of a unicorn, the domain should be an important consideration.

Domain knowledge is distinct from statistics, data analysis, programming and the purely technical areas, and it is easy to see how that is the case. However, business knowledge is often conflated with domain knowledge, perhaps understandably, because these are both vague and interdisciplinary areas. Business knowledge entails some amount of financial knowledge, unit economics models, strategy, people management, and a range of other skills taught in business schools, and more commonly, learned in organisations on the job. Domain knowledge, however, is like being a kind of human expert system.

Domain knowledge is a result of the system exploration that humans as system builders naturally do. To be able to formulate intelligent hypotheses, the unique cause-effect chains that are relevant to specific systems can be studied and understood. Do humans learn about systems in ways that are different from how machines might explore them, if we were to give them infinite data and computational capability? That is a hard question to answer in this context, and perhaps represents a red herring of sorts. What is

SAITEJA XEROX

Scanned with CamScanner

useful to note, however, is that machine learning models still rely on human-formulated hypotheses. There is the odd example of an expert system that has formulated hypotheses and proved them (as is happening in medicine, these days), but these examples are hardly possible without human intervention.

Now that we have established that human intervention has become necessary in machine learning systems, data science can be seen as a field that relies uniquely on human-formulated hypotheses. While computational power and statistical models help us explore and construct hypotheses, the decisions that are made from this data – that help us define hypotheses, model the data to test these hypotheses, construct mathematical or statistical models of these data, and then evaluate the results of those tests – all of these activities take place with human intervention.

So where does domain fit in? Domain experts are those who have significant experience learning about one or a few interconnected systems in intimate ways. Their ability to develop a gut feel for the system's performance and characteristics helps them leap frog the formulation of hypotheses, and this is their biggest benefit, compared to domain-agnostic data scientists, who merely have the programming, statistics, business and communication skills required to make serious analysis happen.

DOMAIN EXPERTISE AND ANALYSIS PARALYSIS

Domain expertise is probably one fine way to fight off the analysis-paralysis problem that plagues many data science teams. Some data science teams take up significant time and resources to experiment with ideas vastly, and the availability of high performance computing power on tap makes them take hypothesis formulation less seriously. Adversity is truly the mother of inventiveness, and it is, for example, when computing power was at a premium, that some of the most efficient sorting algorithms were devised. Similarly, the availability of computing power and statistical modeling capabilities on a massive scale de-incentivize the need to ask pertinent questions.

Pertinent questions and specific answers lead to tangible decisions and related business improvements. Without the benefit of domain knowledge, this is not possible. Analysis paralysis is a very real phenomenon. Data scientists are susceptible in organizations that value domain expertise, and don't value analytical solutions. In situations where analytical solutions and problem solving are valued, data science that fly blind toting algorithms and machine learning won't come out on top either – they're more likely to hurt the credibility of the data science exercise than help it, when they solve simple problems that have pre-existing domain formulations with the help of complex algorithms (which may sometimes not give sufficient insight into their own workings, despite working well).

SAITEJA XEROX

Scanned with CamScanner

CHALLENGE OR CHANNEL DOMAIN EXPERTISE?

Machine learning work done in medicine (cancer cell detection) points to a future where human-learned skills are replicated by deep learning or reinforcement learning systems. Alternatively, many real data science programs at diverse companies indicate an analysis paralysis that can be addressed by involvement to a greater degree of domain experts of specific kinds in the data science hypothesis formulation, analysis and interpretation of results. The latter is more representative of a real world scenario than the former, where an expert system independently learns about a hard problem and solves it.

DOING DATA SCIENCE BETTER

In order to be able to do data science better, it isn't merely important to consider developing data scientist resources along the lines described by Drew Conway or Stephan Kolassa. It is important to groom analytically capable people from within domains too. This means distributing the skill set required for serious analysis from the mainstream data science practice, into functional teams. Sometimes, this may mean penetrating leadership teams that work in functional capacities, and at other times, it may mean addressing the needs of small teams directly, by grooming functional/technical talent for doing data science.

Doing data science better doesn't merely involve leveraging algorithms and their strengths better. It also means asking the right questions.

Place for imagination

Integrating data and design is not new. Budget cuts and decreasing demand during the Great Depression forced brands to improve the effectiveness of mass-market advertisements. Today, the opportunities to design for the individual are overwhelming and amplified by the accelerated access to machine learning. Consumer insights have never looked so plentiful, but the challenge remains how you find and act on them.

The proliferation of digital channels today presents an additional challenge: now more than ever, consumers expect moment-to-moment recognition of their interactions with our brand, including their likes and preferences. We must integrate customer intelligence teams with imaginative creative that brings to life a moment, or concept, in a meaningful and seamless way for the consumer. Creative directors will continue to attempt to capture lightning in a bottle, but a successful holistic strategy depends on executing empathetic designs tailored to individual needs. Machine learning enables individual customer experiences at every moment of the day, keeping customers happy and converting.

SAITEJA XEROX

Scanned with CamScanner

Going to the heart of it, it appears that the true formula lies at the intersection of a unified view of each customer and a data science approach that identifies an actionable insight. With data increasingly commoditized, yet provided freely by consumers, machine learning is the new currency, leading organizations rich with AI.

So, what does a data-rich and creatively driven interaction look like when defined as intelligent and imaginative machine learning at scale? Let's call this intelligent imagination. Here are four actions to consider on this journey:

Create clear and cross functionally aligned marketing objectives that drive a specific outcome

Cultivate imaginative problem solving in your intelligence teams, and data-driven designs in your creative teams, through setting objectives and goals that bridge both functions. We should focus on pressing customer experiences that are specific to the functional team developing them, while tying each experience to a clear business objective.

Facilitate brainstorming between intelligence and imagination functions

We must develop cross-functional processes that take consumer insight from intelligence functions and use this to collaborate among intelligence, strategic and creative teams. Traditionally, the intelligence function may recommend a hypothesis with a loss of interpretation when handing off between teams. This human-led process can be improved through live cross-functional exchanges that refine ideas through an evaluation of creativity, discovery and feasibility while solutions are in their infancy. Collaboration between intelligence and imagination functions also encourages each function to design solutions creatively and holistically through repeated exposure. For example, Spotify has done this well by tapping into listener data and humanizing tech through its hyperlocal out-of-home billboards and advertising campaigns.

Observe people, behavior and content while integrating marketing infrastructure to deliver 1:1 at scale

Supporting technology infrastructure must enable fast information transference from consumer interactions with your brand—in any channel—so that it can be integrated, layered and acted upon in real time. But this alone is not enough. We must recognize that data processed today is a far more powerful predictor of behavior tomorrow. A live triangulation between changing people, behavior and content interactions provides richer contextual depth compared with data models that assume tomorrow will replicate behavior that is months or years old. Such depth allows for more accurate assumptions about likes and interests.

SAI TEJA XEROX

Scanned with CamScanner

Delivery company Postmates does a nice job with using real-time data to optimize campaigns that inform the present and the future. By utilizing a steady stream of data into its tech stacks, the company conceptually orientates all data points around customers and their touch points.

Your moment-to-moment outcomes must be measurable and used to inform future experiences

Your data platform should enable you to understand the constellation of points that match human behavior and experience across their lifecycle. All of these outputs should be measured, but it shouldn't stop there. Your system should continually create and test new hypotheses because behavior, content and inventory are all transient and your algorithms must automatically digest this change. The pattern of taking data offline, testing it and trying to replicate it in production leads to old insights that will not lead to human results.

Ultimately, the winning formula must always be backed by the power of intelligent imagination. Machine learning is at the seat of intelligent imagination, and a sophisticated yet often subtle understanding of needs will yield long-term gains.

Feature Selection algorithms

A feature selection algorithm can be seen as the combination of a search technique for proposing new feature subsets, along with an evaluation measure which scores the different feature subsets. The simplest algorithm is to test each possible subset of features finding the one which minimizes the error rate. This is an exhaustive search of the space, and is computationally intractable for all but the smallest of feature sets. The choice of evaluation metric heavily influences the algorithm, and it is these evaluation metrics which distinguish between the three main categories of feature selection algorithms: wrappers, filters and embedded methods.

Filter Methods

Filter methods are generally used as a preprocessing step. The selection of features is independent of any machine learning algorithms. Instead, features are selected on the basis of their scores in various statistical tests for their correlation with the outcome variable. The correlation is a subjective term here. For basic guidance, you can refer to the following table for defining correlation co-efficients.

Feature\Response	Continuous	Categorical
Continuous	Pearson's Correlation	LDA
Categorical	Anova	Chi-Square

Scanned with CamScanner

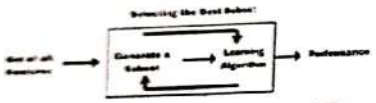
- **Pearson's Correlation:** It is used as a measure for quantifying linear dependence between two continuous variables X and Y. Its value varies from -1 to +1. Pearson's correlation is given as:

$$\rho_{X,Y} = \frac{\mathrm{cov}(X,Y)}{\sigma_X \sigma_Y}$$

- **LDA:** Linear discriminant analysis is used to find a linear combination of features that characterizes or separates two or more classes (or levels) of a categorical variable.
- **ANOVA:** ANOVA stands for Analysis of variance. It is similar to LDA except for the fact that it is operated using one or more categorical independent features and one continuous dependent feature. It provides a statistical test of whether the means of several groups are equal or not.
- **Chi-Square:** It is a is a statistical test applied to the groups of categorical features to evaluate the likelihood of correlation or association between them using their frequency distribution.

One thing that should be kept in mind is that filter methods do not remove multicollinearity. So, you must deal with multicollinearity of features as well before training models for your data.

Wrapper Methods



In wrapper methods, we try to use a subset of features and train a model using them. Based on the inferences that we draw from the previous model, we decide to add or remove features from your subset. The problem is essentially reduced to a search problem. These methods are usually computationally very expensive.

Some common examples of wrapper methods are forward feature selection, backward feature elimination, recursive feature elimination, etc.

- **Forward Selection:** Forward selection is an iterative method in which we start with having no feature in the model. In each iteration, we keep adding the feature which best improves our model till an addition of a new variable does not improve the performance of the model.
- **Backward Elimination:** In backward elimination, we start with all the features and removes the least significant feature at each iteration which improves the performance of the model. We repeat this until no improvement is observed on removal of features.

SAITHJA XEROX

Scanned with CamScanner

- **Recursive Feature elimination:** It is a greedy optimization algorithm which aims to find the best performing feature subset. It repeatedly creates models and keeps aside the best or the worst performing feature at each iteration. It constructs the next model with the left features until all the features are exhausted. It then ranks the features based on the order of their elimination.

One of the best ways for implementing feature selection with wrapper methods is to use Boruta package that finds the importance of a feature by creating shadow features.

It works in the following steps:

1. Firstly, it adds randomness to the given data set by creating shuffled copies of all features (which are called shadow features).
2. Then, it trains a random forest classifier on the extended data set and applies a feature importance measure (the default is Mean Decrease Accuracy) to evaluate the importance of each feature where higher means more important.
3. At every iteration, it checks whether a real feature has a higher importance than the best of its shadow features (i.e. whether the feature has a higher Z-score than the maximum Z-score of its shadow features) and constantly removes features which are deemed highly unimportant.
4. Finally, the algorithm stops either when all features get confirmed or rejected or it reaches a specified limit of random forest runs.

Decision Trees

A **decision tree is** a decision support tool that uses a tree-like graph or model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility. It is one way to display an algorithm that only contains conditional control statements.

Decision trees are commonly used in operations research, specifically in decision analysis, to help identify a strategy most likely to reach a goal, but are also a popular tool in machine learning.

A decision tree is a flowchart-like structure in which each internal node represents a "test" on an attribute (e.g. whether a coin flip comes up heads or tails), each branch represents the outcome of the test, and each leaf node represents a class label (decision taken after computing all attributes). The paths from root to leaf represent classification rules.

In decision analysis, a decision tree and the closely related influence diagram are used as a visual and analytical decision support tool, where the expected values (or expected utility) of competing alternatives are calculated.

SAITEJA XEROX

Scanned with CamScanner

A decision tree consists of three types of nodes:

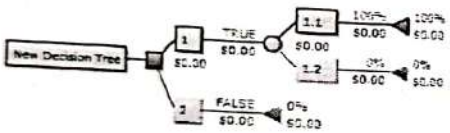
1. Decision nodes – typically represented by squares
2. Chance nodes – typically represented by circles
3. End nodes – typically represented by triangles

Decision trees are commonly used in operations research and operations management. If, in practice, decisions have to be taken online with no recall under incomplete knowledge, a decision tree should be paralleled by a probability model as a best choice model or online selection model algorithm. Another use of decision trees is as a descriptive means for calculating conditional probabilities.

Decision trees, influence diagrams, utility functions, and other decision analysis tools and methods are taught to undergraduate students in schools of business, health economics, and public health, and are examples of operations research or management sciencemethods.

Decision tree building blocks

Decision tree elements



Drawn from left to right, a decision tree has only burst nodes (splitting paths) but no sink nodes (converging paths). Therefore, used manually, they can grow very big and are then often hard to draw fully by hand. Traditionally, decision trees have been created manually — as the aside example shows — although increasingly, specialized software is employed.

Decision rules

The decision tree can be linearized into **decision rules**, where the outcome is the contents of the leaf node, and the conditions along the path form a conjunction in the if clause. In general, the rules have the form:

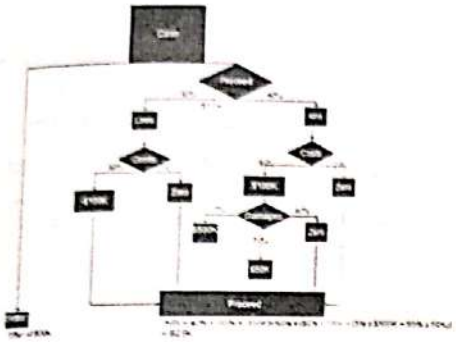
if condition1 *and* condition2 *and* condition3 *then* outcome.

Decision rules can be generated by constructing association rules with the target variable on the right. They can also denote temporal or causal relations.

Scanned with CamScanner

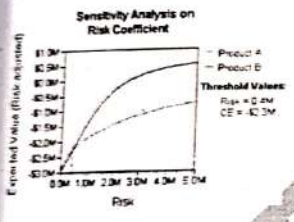
Decision tree using flowchart symbols

Commonly a decision tree is drawn using flowchart symbols as it is easier for many to read and understand.



Analysis example

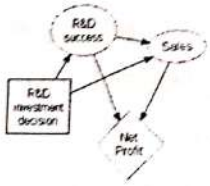
Analysis can take into account the decision maker's (e.g., the company's) preference or utility function. for example:



The basic interpretation in this situation is that the company prefers B's risk and payoffs under realistic risk preference coefficients (greater than $400K—in that range of risk aversion, the company would need to model a third strategy, "Neither A nor B").

Influence diagram

Much of the information in a decision tree can be represented more compactly as an influence diagram, focusing attention on the issues and relationships between events.



The rectangle on the left represents a decision, the ovals represent actions, and the diamond represents results.

SAITEJA XEROX

Scanned with CamScanner

Association rule induction

Decision trees can also be seen as generative models of induction rules from empirical data. An optimal decision tree is then defined as a tree that accounts for most of the data, while minimizing the number of levels (or "questions").

Random forests

Random forests or **random decision forests** are an ensemble learning method for classification, regression and other tasks, that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. Random decision forests correct for decision trees' habit of overfitting to their training set.

The first algorithm for random decision forests was created by Tin Kam Housing the random subspace method, which, in Ho's formulation, is a way to implement the "stochastic discrimination" approach to classification proposed by Eugene Kleinberg.

An extension of the algorithm was developed by Leo Breiman and Adele Cutler, and "Random Forests" is their trademark. The extension combines Breiman's "bagging" idea and random selection of features, introduced first by Ho and later independently by Amit and Geman in order to construct a collection of decision trees with controlled variance.

Algorithm

Preliminaries: decision tree learning

Decision trees are a popular method for various machine learning tasks. Tree learning "come[s] closest to meeting the requirements for serving as an off-the-shelf procedure for data mining", say Hastie *et al.*, "because it is invariant under scaling and various other transformations of feature values, is robust to inclusion of irrelevant features, and produces inspectable models. However, they are seldom accurate".

In particular, trees that are grown very deep tend to learn highly irregular patterns: they overfit their training sets, i.e. have low bias, but very high variance. Random forests are a way of averaging multiple deep decision trees, trained on different parts of the same training set, with the goal of reducing the variance. This comes at the expense of a small increase in the bias and some loss of interpretability, but generally greatly boosts the performance in the final model.

SAITEJA XEROX

Scanned with CamScanner

Tree bagging

The training algorithm for random forests applies the general technique of bootstrap aggregating, or bagging, to tree learners. Given a training set $X = x_1, \ldots, x_n$ with responses $Y = y_1, \ldots, y_n$ bagging repeatedly (B times) selects a random sample with replacement of the training set and fits trees to these samples:

For $b = 1, \ldots, B$:

1. Sample, with replacement, n training examples from X, Y; call these X_b, Y_b.
2. Train a classification or regression tree f_b on X_b, Y_b.

After training, predictions for unseen samples x' can be made by averaging the predictions from all the individual regression trees on x':

$$\hat{f} = \frac{1}{B}\sum_{b=1}^{B} f_b(x')$$

or by taking the majority vote in the case of classification trees.

This bootstrapping procedure leads to better model performance because it decreases the variance of the model, without increasing the bias. This means that while the predictions of a single tree are highly sensitive to noise in its training set, the average of many trees is not, as long as the trees are not correlated. Simply training many trees on a single training set would give strongly correlated trees (or even the same tree many times, if the training algorithm is deterministic); bootstrap sampling is a way of de-correlating the trees by showing them different training sets.

Additionally, an estimate of the uncertainty of the prediction can be made as the standard deviation of the predictions from all the individual regression trees on x':

$$\sigma = \sqrt{\frac{\sum_{b=1}^{B}(f_b(x') - \hat{f})^2}{B-1}}.$$

The number of samples/trees, B, is a free parameter. Typically, a few hundred to several thousand trees are used, depending on the size and nature of the training set. An optimal number of trees B can be found using cross-validation, or by observing the *out-of-bag error*: the mean prediction error on each training sample x_i, using only the trees that did not have x_i in their bootstrap sample. The training and test error tend to level off after some number of trees have been fit.

From bagging to random forests

The above procedure describes the original bagging algorithm for trees. Random forests differ in only one way from this general scheme: they use a modified tree learning algorithm that selects, at each candidate split in the learning process, a random subset of the features. This process is sometimes called "feature bagging". The reason for doing this is the correlation of the trees in an ordinary bootstrap sample: if one or a few features are very strong predictors for the response variable (target output), these features will be selected in many of the B trees, causing them to become correlated. An analysis of how bagging and random subspace projection contribute to accuracy gains under different conditions is given by Ho.

Typically, for a classification problem with p features, $\sqrt{p}$ (rounded down) features are used in each split. For regression problems the inventors recommend $p/3$ (rounded down) with a minimum node size of 5 as the default.

ExtraTrees

Adding one further step of randomization yields *extremely randomized trees*, or ExtraTrees. These are trained using bagging and the random subspace method, like in an ordinary random forest, but additionally the top-down splitting in the tree learner is randomized. Instead of computing the locally *optimal* feature/split combination (based on, e.g., information gain or the Gini impurity), for each feature under consideration, a random value is selected for the split. This value is selected from the feature's empirical range (in the tree's training set, i.e., the bootstrap sample).

Scanned with CamScanner

UNIT- IV

Data Visualization

Data visualisation is viewed by many disciplines as a modern equivalent of visual communication. It involves the creation and study of the visual representation of data, meaning "information that has been abstracted in some schematic form, including attributes or variables for the units of information".

A primary goal of data visualization is to communicate information clearly and efficiently via statistical graphics, plots and information graphics. Numerical data may be encoded using dots, lines, or bars, to visually communicate a quantitative message. Effective visualization helps users analyze and reason about data and evidence. It makes complex data more accessible, understandable and usable. Users may have particular analytical tasks, such as making comparisons or understanding causality, and the design principle of the graphic (i.e., showing comparisons or showing causality) follows the task. Tables are generally used where users will look up a specific measurement, while charts of various types are used to show patterns or relationships in the data for one or more variables.

Data visualization is both an art and a science. It is viewed as a branch of descriptive statistics by some, but also as a grounded theory development tool by others. Increased amounts of data created by Internet activity and an expanding number of sensors in the environment are referred to as "big data" or Internet of things. Processing, analyzing and communicating this data present ethical and analytical challenges for data visualization. The field of data science and practitioners called data scientists help address this challenge.

Basic principles

We begin with some basic ideas about data visualization from Edward Tufte (The Visual Display of Quantitative Information

He gives six principles of graphical integrity

The first two are

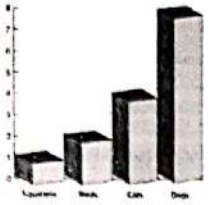
1. The representation of numbers, as physically measured on the surface of the graphic itself, should be directly proportional to the numerical quantities measured.

2. Clear, detailed, and thorough labeling should be used to defeat graphical distortion and ambiguity. Write out explanations of the data on the graphic itself. Label important events in the data.

SAI TEJA XEROX

Scanned with CamScanner

The classic example to avoid: adding 3d where it isn't necessary



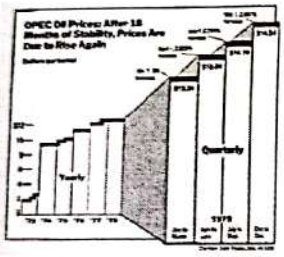
"The Lie Factor" is defined as:

$$\text{lie factor} = \frac{\text{size of effect shown in graphic}}{\text{size of effect in data}}$$

When this isn't 1, it is not uncommon to be as large as 2 to 5

It is less usual to see a lie factor of less than 1

The third principle of graphical integrity: Show data variation, not design variation For example, avoid things like this:



The fourth principle of graphical integrity: In time-series displays of money, deflated and standardized units of monetary measurement are nearly always better than nominal units.

The fifth : The number of information-carrying (variable) dimensions depicted should not exceed the number of dimensions in the data

Tufte's sixth principle of graphical integrity is: Graphics must not quote data out of context.

To evaluate the information content, a useful measure is:

$$\text{Data ink ratio} = \frac{\text{data ink}}{\text{total ink used in the graphic}}$$

This is the same as: the proportion of a graphic's ink devoted to the non-redundant display of data information

It is also the same as: 1 minus the proportion of a graphic that can be erased without the loss of data information

One of Tufte's principles for graphical excellence: Maximize the data ink ratio, within reason.

SAITEJA XEROX

Scanned with CamScanner

This is why Tufte frowns on heavy grid lines in the background (or even horizontal reference lines)

He extends this principle with two guidelines for how to achieve it

1. Erase non-data ink, within reason.

2. Erase redundant data ink, within reason.

To summarize Tufte's principles of how to achieve graphical excellence:

1. Above all else show the data.

2. Maximize the data-ink ratio.

3. Erase non-data ink.

4. Erase redundant data ink.

5. Revise and edit.

A useful term coined by Tufte: "chartjunk refers to all visual elements in charts and graphs that are not necessary to comprehend the information represented on the graph, or that distract the viewer from this information" (quoted from Wikipedia)

Common types of chartjunk:

1. Vibrating chartjunk, which is cross-hatching or other patterns that distract the mind from the information being presented

2. Grids (according to Tufte) — make them a lighter gray, not black, if you are going to use them

3. Self-promoting graphics ("The Duck"), where color schemes and patterns are introduced for artistic appeal rather than information content.

Tufte gives one other useful pair of principles on the aspect ratio of a display:

- If the nature of the data suggests the shape of the graphic, follow that suggestion.
- Otherwise, move toward horizontal graphics about 50 percent wider than tall (approximately a 3:2 aspect ratio).

We now describe some principles given by William S. Cleveland in his book *The Elements of Graphing Data* (1985)

Many (but not all) of these are quite similar to those of Tufte, although his work in this area preceded Tufte's work

His four major categories in the principles of graph construction:

SAITEJA XEROX

Scanned with CamScanner

- Clear vision
- Clear understanding
- Scales
- General strategy

Clear vision

- Make the data stand out. Avoid superfluidity.
- Use visually prominent graphical elements to show the data.
- Do not clutter the data region.
- Use a reference line when there is an important value that must be seen across the entire graph, but do not let the line interfere with the data.
- Do not allow data labels in the data region to interfere with the quantitative data or to clutter the graph.
- Avoid putting notes, keys, and markers in the data region. Put keys and markers just outside the data region and put notes in the legend or in the text.
- Overlapping plotting symbols must be visually distinguishable.
- Superposed data sets must be readily visually discriminated.
- Visual clarity must be preserved under reduction and reproduction.

Clear understanding

- Put major conclusions into graphical form. Make legends comprehensive and informative.
- Error bars should be clearly explained.
- Proofread graphs.
- Strive for clarity.

Scales

- Choose the range of the tick marks to include or nearly include the range of data.
- Subject to the constraints that scales have, choose the scales so that the data fill up as much of the region as possible.
- It is sometimes helpful to use the pair of scale lines for a variable to show two different scales.
- Choose appropriate scales when graphs are compared.
- Do not insist that zero always be included on a scale showing magnitude.
- Use a logarithmic scale when it is important to understand percent change or multiplicative factors.
- Showing data on a logarithmic scale can improve resolution.

SAI TEJA XEROX

Scanned with CamScanner

General strategy

- A large amount of quantitative information can be packed into a small region.
- Graphing data should be an iterative, experimental process.
- Graph data two or more times when it is needed.
- Many useful graphs require careful, detailed study

Ideas and tools for data visualization

All the data in the world is useless – in fact it can become a liability – if you can't understand it. Data visualization is about how to present your data, to the right people, at the right time, in order to enable them to gain insights most effectively.

Luckily visualization solutions are evolving as rapidly as the rest of the tech stack. Charts, videos, infographics and at the cutting edge even virtual reality and augmented reality (VR & AR) presentations offer increasingly engaging and intuitive channels of communication.

Here's some of the best, most popular or most innovative data visualization tools available today.

Tableau

Tableau is often regarded as the grand master of data visualization software and for good reason. Tableau has a very large customer base of 57,000+ accounts across many industries due to its simplicity of use and ability to produce interactive visualizations far beyond those provided by general BI solutions. It is particularly well suited to handling the huge and very fast-changing datasets which are used in Big Data operations, including artificial intelligence and machine learning applications, thanks to integration with a large number of advanced database solutions including Hadoop, Amazon AWS, My SQL, SAP and Teradata. Extensive research and testing has gone into enabling Tableau to create graphics and visualizations as efficiently as possible, and to make them easy for humans to understand.

Qlikview

Qlik with their Qlikview tool is the other major player in this space and Tableau's biggest competitor. The vendor has over 40,000 customer accounts across over 100 countries, and those that use it frequently cite its highly customizable setup and wide feature range as a key advantage. This however can mean that it takes more time to get to grips with and use it to its full potential. In addition to its data visualization capabilities Qlikview offers powerful business intelligence, analytics and enterprise reporting capabilities and I particularly like the clean and clutter-free user interface. Qlikview is commonly used alongside its sister package, Qliksense, which handles data exploration and discovery. There is also a strong

Scanned with CamScanner

community and there are plenty of third-party resources available online to help new users understand how to integrate it in their projects.

FusionCharts

This is a very widely-used, JavaScript-based charting and visualization package that has established itself as one of the leaders in the paid-for market. It can produce 90 different chart types and integrates with a large number of platforms and frameworks giving a great deal of flexibility. One feature that has helped make FusionCharts very popular is that rather than having to start each new visualization from scratch, users can pick from a range of "live" example templates, simply plugging in their own data sources as needed.

Highcharts

Like FusionCharts this also requires a licence for commercial use, although it can be used freely as a trial, non-commercial or for personal use. Its website claims that it is used by 72 of the world's 100 largest companies and it is often chosen when a fast and flexible solution must be rolled out, with a minimum need for specialist data visualization training before it can be put to work. A key to its success has been its focus on cross-browser support, meaning anyone can view and run its interactive visualizations, which is not always true with newer platforms.

Datawrapper

Datawrapper is increasingly becoming a popular choice, particularly among media organizations which frequently use it to create charts and present statistics. It has a simple, clear interface that makes it very easy to upload csv data and create straightforward charts, and also maps, that can quickly be embedded into reports.

Plotly

Plotly enables more complex and sophisticated visualizations, thanks to its integration with analytics-oriented programming languages such as Python, R and Matlab. It is built on top of the open source d3.js visualization libraries for JavaScript, but this commercial package (with a free non-commercial licence available) adds layers of user-friendliness and support as well as inbuilt support for APIs such as Salesforce.

Sisense

Sisense provides a full stack analytics platform but its visualization capabilities provide a simple-to-use drag and drop interface which allow charts and more complex graphics, as well as interactive visualizations, to be created with a minimum of hassle. It enables multiple sources of data to be gathered

SAITHAXBOX

Scanned with CamScanner

into one easily accessed repositories where it can be queried through dashboards instantaneously, even across Big Data-sized sets. Dashboards can then be shared across organizations ensuring even non technically-minded staff can find the answers they need to their problems.

Example of inspiring (industry) projects

The first thing that springs to mind when it comes to **data visualization** is, of course, infographics. This sort of visual representation that, as a rule, includes charts, graphs, illustrations and other graphical stuff handles various issues on a professional level. However boring and dull statistics can be, when it is translated into an infographic it takes a different form.

Much like any other means, it has the ability to react to current tendencies and skillfully adapt to the changes thereby maturing and perfecting. From primitive Isotype-like projects to sophisticated visualization artworks, it has come a long way.

The design trends also affect it. For example, while the skeuomorphism held rein: infographicswere populated with fake 3d objects and elements with shadows or glossy surfaces. In the era of flat style, everything began to look plain, neat and subtle.

Right now is the period of web-based interactive infographics that employ the latest design trends, high-end techniques and even pioneering WebGL features. Some of them are just basic combinations of dynamic charts and animated icons, while others are more advanced projects that professionally deal with the real-time data and actively cooperate with the readers.

Today we are going to take a look at ten matchless data visualization projects.

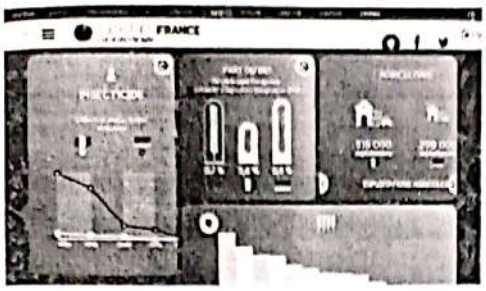
Outstanding Infographics

1 Jour en France

1 Jour en France is an outstanding project that is charged with the top-notch techniques. It features a location-based data visualization that is broken into three main parts. The first one is an enthralling quiz; the second one is a reply video experience, and the third one is an interactive circular dashboard with some essential functionality. The statistics are displayed via a card-centric layout filled with simple yet eye-pleasing graphics.

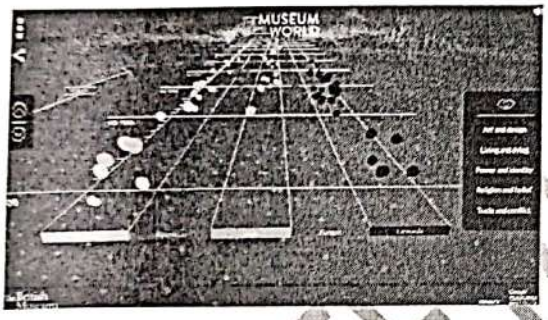
SAI TEJA XEROX

Scanned with CamScanner



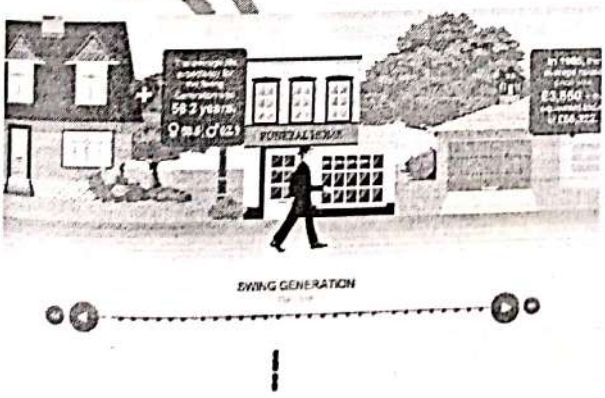
The Museum of the World

When Google is involved, we should expect something incredible and innovative. The Museum of the World is vivid proof of that. The website invites the online audience to join the adventure through history, in a fascinating way. It is an inventively reproduced timeline. It resembles a speedway that is enriched with interesting facts presented as colorful bubbles. There are several tracks that identify various continents. You can use key arrows or just scroll up and down to move around.



WizzCash

The team behind WizzCash has taken into account the latest research and numbers to illustrate the so-called generation debt. They have come up with a small but absorbing trip that leads through the four generations (Swing Generation, Baby Boomers, Generation X and MILLENNIALS). It demonstrates today's economic realities and lets you compare and contrast the situations. Here all the facts are at your fingertips: just follow the characters to find out more.



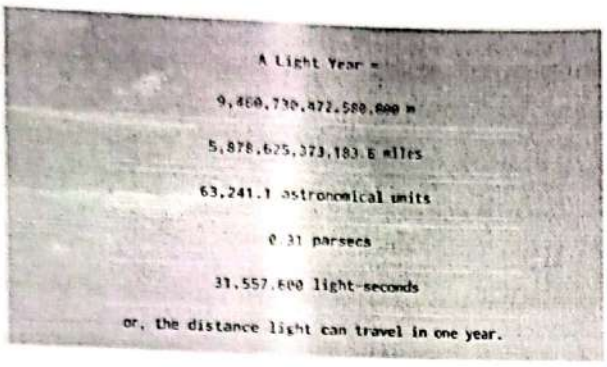
G.PRITHVI M.Com, M.Sc, M.B.A, P.G.D.H.M, P.G.D.G.C, P.G.D.C.A

SAITEJA XEROX

Scanned with CamScanner

Exercise: Create your own Visualization of a complex data set

A Light Year



What is it?

This website is like a day calendar, showcasing the objects that a beam of light might pass by, when traveling from the sun. in a straight line, for one year.

Where is the data from?

Not specified.

UNIT – V

Data Science and Ethical Issues

Those who work in data science and analytics are at the coalface of digital transformation, and opportunities for achieving social good within the field of data science are huge. In just a few short years, data has become one of the most valuable commodities in the global economy.

The vast quantities of user data from social media accounts, apps and IoT devices have created swathes of information from which vital insights can be mined to implement new problem-solving technologies, ostensibly making the world a better place.

That hinges on the continuing development of an ethical framework, increased transparency with the public about their data and constant cognisance of potential biases (both human and within datasets) that could be used in automated decision-making or other technological advances.

Discussions on privacy

If you're not sure whether the data fueling your approach will raise privacy and security flags, consider the following. When it comes to data privacy and security, not all data is going to be of equal concern. Much depends on the level of detail in data content, data type, data structure, volume, and velocity, and

SAITEJA XEROX

Scanned with CamScanner

indeed how the data itself will be used and released.

First there is the data where security and privacy has always mattered and for which there is already an existing and well galvanized body of law in place. Foremost among these is classified or national securitydata where data usage is highly regulated and enforced. Other data for which there exists a considerable body of international and national law regulating usage includes:

- **Proprietary Data** – specifically the data that makes up the intellectual capital of individual businesses and gives them their competitive economic advantage over others, including data protected under copyright, patent, or trade secret laws and the sensitive, protected data that companies collect on behalf of its customers;
- **Infrastructure Data** - data from the physical facilities and systems – such as roads, electrical systems, communications services, etc. – that enable local, regional, national, and international economic activity: and
- **Controlled Technical Data** - technical, biological, chemical, and military-related data and research that could be considered of national interest and be under foreign export restrictions.

It may be possible to work with publicly released annualized and cleansed data within these areas without a problem, but the majority of granular data from which significant insight can be gleaned is protected. In most instances, scientists, researchers, and other authorized developers take years to appropriately acquire the expertise, build the personal relationships, and construct the technical, procedural and legal infrastructure to work with the granular data before implementing any approach. Even using publicly released datasets within these areas can be restricted, requiring either registration, the recognition of or affiliation to an appropriate data governing body, background checks, or all three before authorization is granted.

The second group of data that raises privacy and security concerns is personal data. Commonly referred to as Personally Identifiable Information (PII), it is any data that distinguishes individuals from each other. It is also the data that an increasing number of digital approaches rely on, and the data whose use tends to raise the most public ire. Personal data could include but is not limited to an individual's:

- Government issued record data (social security numbers, national or state identity numbers, passport records, vehicle data, voting records, etc.);
- Law enforcement data (criminal records, legal proceedings, etc.);
- Personal financial, employment, medical, and education data;

SATHIA XEROX

Scanned with CamScanner

- Communication records (phone numbers, texts data, message records, content of conversations, time and location, etc.);
- Travel data (when and where traveling, carriers used, etc.);
- Networks and memberships (family, friends, interests, group affiliations, etc.);
- Location data (where a person is and when);
- Basic contact information (name, address, e-mail, telephone, fax, twitter handles, etc.);
- Internet data (search histories, website visits, click rates, likes, site forwards, comments, etc.);
- Media data (which shows you're watching, music you're listening to, books or magazines you're reading, etc.);
- Transaction data (what you're buying or selling, who you're doing business with, where, etc.); and
- Bio and activity data (from personal mobile and wearable devices).

In industries where being responsible for handling highly detailed personal data is the established business norm – such as in the education, medical and financial fields – there are already government regulations, business practices and data privacy and security laws that protect data from unauthorized usage, including across new digital platforms. But in many other industries, particularly in data driven industries where personal data has been treated as proprietary data and become the foundation of business models, there is currently little to no regulation.

Regulation protecting privacy was already an issue before the digitization of data, when the largely manual process of collecting data through industry databases, surveys, phone books, and public records resulted in junk mail, telemarketing, and annoying robocalls. But now that businesses can automatically collect passive and active personal data via consumer behaviors and multiple physical sensors, consumers have become significantly more alarmed at how customization in services and products can lead to the loss of personal privacy. Take for example the growth in voices raised over the use of data from Wearables or from sensors related to the interconnected "Internet of Things". In the new normal, the more that a data approach depends on data actively or passively collected on individuals, the more likely that consumers will speak up and demand privacy protection, even if they previously gave some form of tacit approval to use their data.

SAITHJA XEROX

Scanned with CamScanner

Despite this new landscape, there are lots of different ways to use personal data, some of which may not trigger significant privacy or security concerns. This is particularly true in cases where individuals willingly provide their data or data cannot be attributed to an individual. Whether individuals remain neutral to data approaches tends to be related to the level of control they feel they have over how their personal data is used. Some organizations that collect personal data extensively, such as Facebook and Google, work to increasingly provide their users with methods to control their own data. But for others, the lack of due diligence on data privacy in their approaches has already had their effect.

A third category of data needing privacy consideration is the data related to good people working in difficult or dangerous places. Activists, journalists, politicians, whistle-blowers, business owners, and others working in contentious areas and conflict zones need secure means to communicate and share data without fear of retribution and personal harm. That there are parts of the world where individuals can be in mortal danger for speaking out is one of the reason that TOR (The Onion Router) has received substantial funding from multiple government and philanthropic groups, even at the high risk of enabling anonymized criminal behavior. Indeed, in the absence of alternate secure networks on which to pass data, many would be in grave danger, including those such as the organizers of the Arab Spring in 2010 as well as dissidents in Syria and elsewhere.

Yet it is easy for well-meaning, idealistic developers to implement worldwide services that do not take data privacy and security considerations into account simply because there was no perceived need for them and no laws, regulations or procedures to guide the process. Take for instance the experiences of the teams of volunteers seeking to build on successful efforts to crowdsource and geo-locate humanitarian crises around the globe. Crowdmapping initiatives have been effective in humanitarian contexts, like relief efforts in Nepal, where everyone is willing to share data for the greater good. But when it comes to geo-locating data on conflict events, spikes of alarm over the veracity of data and the safety of those who would provide good data increase profoundly.

Security

Understanding that big data comes from such a huge pool of devices, constantly collecting new data, gives you an idea where security threats may come from. Each device that is online, from phones to tablets to computers to smart appliances, has the potential of sending and receiving new data. This continuous back and forth gives a nearly constant window of opportunity for a security breach. In order to prevent this from happening, you must be aware of where you could face issues.

SAITEJA XEROX

Scanned with CamScanner

1. Data Distribution Over Multiple Frameworks

While taking that vast amount of data and processing and distributing all of it over multiple different frameworks helps to keep the data moving more efficiently, it also opens the data up to larger headaches. Each server that handles the big data coming in has the potential of falling to a network security breach or getting locked up in the network as the process happens. The speed of data may be faster, but the multiple entry points to the same pot of information leaves it more vulnerable than if a single framework is used.

2. Real Time Network Security and Compliance Tools

It is great to monitor your network. It helps to keep your network a lot safer. However, these tools are also stockpiles of massive amounts of big data. They have to search through numerous bits of data while constantly making sure your network security is properly in place. Occasionally, that means getting a false-positive result in what appears to be an outside intrusion. In order to keep these systems safe, they need to be checked by humans to keep only the actual threats being tended to. Using a real time network monitor tool gives your business the ability to sift through the data coming in, reducing the false positives and giving your network security tools the best chances of catching real intrusions as they occur.

3. Data Storage

When big data gets stored, it is sitting there, waiting to be accessed. If you are the one accessing it, then there is no problem. However, if an outsider is accessing the data while it is stored, this type of breach may be missed. Since stored data is rarely monitored for someone accessing it, it could be a virtual goldmine for someone trying to sneak their way into your network.

4. Network Access Points

One of the ways that your network can keep big data safer for you is by doing everything you can to protect the access points of your data. You need to make sure everything you send along is encrypted. You also need to make sure that you verify each person is exactly who they claim to be. These should be multi-stage verification so that you can tell who is looking at what, and that they are only seeing the data they are allowed to see. Anything beyond what they are allowed to see needs to be kept safe and not put out there for anyone to see.

SAITEJA XEROX

Scanned with CamScanner

5. Non-Related Stores of Data

Having data that is stored with yours, but is not related to yours, could pose another potential issue when it comes to keeping your big data safe. Consider a NoSQL database. They are typically only kept secure with a bit of middle-ware, but that tends to be a weak level of network security. You need to have more protection than someone else's software that is not typically strong enough to keep those determined to get in, out.

6. Data Endpoints

When the data is sent along to be stored, it sits at its endpoint and waits for you to do something with it. If your endpoint is not secure, you could have issues with the data being stolen or corrupted while it sits there. If you have a network security breach along the way, your data could be stopping at what someone else has told the data is the endpoint, and not even getting to where you want your data to be. You need to verify the data is going where you want it to go and that it is safe there on a regular basis to avoid this type of issue.

7. Data Mining Solutions

Data mining is what gives you access to the answers that all of this big data collecting is looking for. It helps you find new trends that people are leaning towards, and helps you to develop business strategies that take those new trends into account. That is the reason you need to make sure you have top level network security monitoring this information. Not only outsiders are going to think that this information is valuable, as insiders could be seeking a way to be a part of the next big thing, too. Making sure your network security is all inclusive is the best way to ensure that these results are kept safe until you can do what you need to with them.

8. Audits That Find Data Breaches Have Occurred

When you are talking about loss of big data, when you find that someone missed an intrusion, that is already a loss of the data, leaving it exposed to issues. These audits need to show you what happened and when, plus what information may have been compromised in the process. While this is a data loss, it is also an experience that allows you to better prepare for future intrusions, and it can help keep your future big data safer.

SAITEJA XEROX

Scanned with CamScanner

9. Metadata Can Give Insight

While the metadata is very important, because it can tell you so much about your data. if you are not the only one seeing your metadata, that is a problem. You can see who accessed your specific data, when it was accessed, and what of the data was accessed. However, people who have access to this can become a breach in and of themselves. If you are going to allow anyone to access your metadata. you need to make sure that you trust them with all of the insider data that your company has. or else. you risk losing your business' information to someone else.

Ethics

Informed consent

The ethical issue of consent arises because in big data analytics, very little may be known about intended future uses of data when it is collected. With such uncertainty. neither benefits nor risks can be meaningfully understood. Thus, it is unlikely that consent obtained at the point of data collection (one-off) would meet a strict definition of "informed consent". For example, procedures exist for "broad" and "generic" consent to share genomic data, but are criticised on the grounds that such consent cannot be meaningful in light of risks of unknown future genetic technologies. In 2002, O'Neill noted how this limitation of consent is not new, but the use of data for such different purposes, and the scale of possible harms make it more problematic with big data.

Even if such conceptual issues are minimised (or ignored), practical challenges remain.

- Obtaining informed consent may be impossible or prohibitively costly due to factors such as scale, or the inability to privately contact data subjects.
- The validity of consent obtained by agreement to terms and conditions is debateable, especially when agreement is mandatory to access a service.

De-identification

Unfortunately, there exist no robust, unanimously internationally agreed definitions for the terms de-identification, anonymisation, and pseudonymisation. Generally, a dataset is said to be de-identified if elements that might immediately identify a person or organisation have been removed or masked. In part because a number of relevant laws, such as data protection legislation, define different treatments for identifiable and nonidentifiable data, much has rested on being able to make this distinction. Despite this legal situation, recognition is growing that such distinctions are becoming less tenable.

SATTEJA XEROX

Scanned with CamScanner

• Identifiability is increasingly being seen as a continuum, not binary.

• Disclosure risks increase with dimensionality (i.e., number of variables), linkage of multiple data sources, and the power of data analytics.

• Disclosure risks can be mitigated, but not eliminated.

• De-identification remains a vital tool to lower disclosure risk, as part of a broader approach to ensuring safe use of data.

Inequality – digital divide

While the benefits of scale in many domains are clear (e.g., medical care), some see risks in the accumulation of data at a new scale with power that entails, whether data is held in public or private institutions. For reasons of scale and complexity, a relatively small number of entities have the infrastructures and skills to acquire, hold, process and benefit from big data.

• While the question of who owns data is a legal one, the consequences of inequality pose ethical questions.

• Who can access data? In principle, any researcher can access Twitter via its API, but the costs and skills needed do present access barriers.

• Who governs data access? Increasingly, data with disclosure risks can be safely curated, with access enabled through governance mechanisms, such as committees. Is such access genuinely equally open? How is this documented?

Research integrity

Data repositories play a vital role in supporting research integrity by holding data and making them available to others for both validation, replication, as well as providing new research opportunities. To do so, data must have clear "provenance", its sources and processing need to be known, identified, and documented. The attenuated relationship between data curators and data producers, who may not be 'researchers' per se, makes this challenging for a number of reasons:

• Much data not collected for research, such as administrative data, has different standards (e.g., quality, metadata) to research data.

• For some genres, often with commercial value, such as Twitter data, there are legal restrictions on reproducing data, including providing data to support publications. For a comprehensive treatment of issues of preserving social media, see Thomson 2016.

G.PRITHVI M.Com, M.Sc, M.B.A, P.G.D.H.M, P.G.D.G.C, P.G.D.C.A

SAI TEJA XEROX

Scanned with CamScanner

• Data repositories face challenges in upholding their commitments to standards oftransparency and reproducibility when working with groups of data producers who do not routinely generate data for social research.

Emerging issues

• Alternatives to individual informed consent, e.g., "social consent" are being tested whereby sufficient protections are in place to ethically permit data use without individual informed consent.

• There is growing recognition of the need to respect the source and provenance of data—and more broadly its "contextual integrity"—when deciding what, if any, reuse is permissible.

• Most research ethics are based on the assumption that the entity at risk is an individual, hence de-identification offers protection. If harms can be inflicted, for example, denial of health care, based on group membership with no need for individual identification, then the protection of de-identification is no longer adequate.

• If it is no longer possible to neatly divide public and private, then some suggest assessing data use based on outcomes, and permitted uses with "public benefit" or in the "public interest". However, definitions are often vague, and such benefits accrue long after the decision about data use has been made. How can data users be held accountable for delivering the promised public well-being?

A look back at Data Science

There's a part of data science that you rarely hear about: the deployment and production of data flows. Everybody talks about how to build models, but little time is spent discussing the difficulties of actually using those models. Yet these production issues are the reason many companies fail to see value come from their data science efforts and investments.

The data science process is extensively covered by resources all over the web and known by everyone. A data scientist connects to data, splits it or merges it, cleans it, builds features, trains a model, deploys it to assess performance, and iterates until they're happy with it. That's not the end of the story though. Next, you need to try the model on real data and enter the production environment.

These two environments are inherently different because the production environment is continuously running – and potentially impacting existing internal or external systems. Data is constantly coming in, being processed and computed into KPIs, and going through models that are retrained frequently. These systems, more often than not, are written in different languages than the data science environment.

To better understand the challenges companies face when taking data science from prototype to production, Dataiku recently asked thousands of companies around the world how they do it. The results

Scanned with CamScanner

show that companies using data science have unique challenges that fall into four different profiles-that they've coined as follows: Small Data Teams, Packagers, Industrialisation Maniacs. and The Big Data Lab.

SMALL DATA TEAMS (23%)

Small Data Teams Focus on building small projects fast: standard machine learning packages with a unique server and technical environment for all analytics projects.

> 3/4 Do either Marketing or reporting.

> 61% Report having custom machine learning as part of their business model.

> 83% Use either SQL or Enterprise Analytics databases.

These teams, as their name indicate, use mostly small data and have a unique design /production environment. They deploy small continuous iterations and have little to no rollback strategy. They often don't retrain models and use simple batch production deployment, with few packages. Business teams are fairly involved throughout the data project design and deployment.

Average level of difficulty of deployment: 6.4

PACKAGERS (27%)

Packagers Focus on Building a Framework (the software development approach): independent teams that build their own framework for a comprehensive understanding of the project.

> 48% have set-up Advanced Reporting.

> 52% of respondents mix storage technologies.

> 63% use SQL and open source.

These teams have a software development approach to data science and have often built their framework from scratch. They develop ad-hoc packaging and practice informal A/B testing. They use Git intensely to understand the globality of their projects and their dependencies, and they are particularly interested in IT environment consistency. They tend to have a multilanguage environment and are often disconnected from business teams.

Average level of difficulty in deployment: 6.4

INDUSTRIALISATION MANIACS (18%)

Industrialisation Maniacs Focus on Versioning and Auditing: IT-driven teams that think in terms of frequent deployment and constant logging to track all changes and dependencies.

G.PRITHVI M.Com, M.Sc, M.B.A, P.G.D.H.M, P.G.D.G.C, P.G.D.C.A

SAITEJA XEROX

Scanned with CamScanner

> 61% have Logistics, Security, or Industry Specific use cases .

> 30% have deployed Advanced Reporting (vs 50% of all respondents)

> 72 % use NoSQL and Cloud.

These data teams are mostly IT-led and don't have a distinct production environment. They have complex automated processes in place for deployment and maintenance. They log all data access and modification and have a philosophy of keeping track of everything. In these setups, business teams are notably not involved in the data science process and monitoring.

Average level of difficulty in deployment: 6.9

THE BIG DATA LAB (30%)

The Big Data Lab Focus' on Governance and Project Management: Mature teams with a global deployment strategy, rollback processes, and preoccupation with governance principles and integration within the company.

>66% of companies have multiple use cases in place.

> 50% do advanced Social Media Analytics (vs 22% of global respondents).

> 53% use Hadoop and two thirds of them only use Hadoop.

These teams are very mature with more complex use cases and technologies. They used advanced techniques such as PMML, multivariate testing (or at least formal A/B testing), have automated procedures to backtest, and robust strategies to audit IT environment consistency. In these larger, more organized teams, business users are extremely involved before and after the deployment of the data product.

Average level of difficulty in deployment: 5.6.

Overall, the main reported barrier to production for all groups (50% of respondents) is data quality and pipeline development issues. In terms of the overall difficulty of data science production, the average reported difficulty of deploying a data project into production is 6.18 out of ten, and 50% of respondents' state that on a scale of 1 to 10, the level of difficulty involved in getting a data product in production is between six and 10.

Considering the results, these are a few principles that companies should keep in mind on how to build production-ready data science products:

1. Getting started is tough. Working with small data on SQL databases does not mean it's going to be easier to deploy into production.

SAI TEJA XEROX

Scanned with CamScanner

2. Multi language environments are not harder to maintain in production, as long as you have an IT environment consistency process. So mix'n'match!

3. Real-time scoring and online machine learning are likely to make your production pie more complex. Think about whether the improvement to your project is worth the hassle.

4. Working with business users, both while designing your machine learning project and after when monitoring it day to day, will increase your efficiency. Collaborate!

Next – generation data scientists

Data science can be a superpower. While we can't solve all the ills of the world with machine learning, applying data science to just about any problem nudges us a little closer to a solution.

The data business is booming right now.The fact that 90% of the data in world today has been created in the last two years is a great example of the growth trajectory of data. All of this data provides new opportunities for discovery for those who are willing to analyze it. Enter the data scientist.

With a growth pattern similar to that of data itself, it's safe to say that data scientists are going to be in high demand. Among other skills, being a practitioner of data science requires analytical thinking, mathematical/statistical ability, a knack for communicating results to non-data people, and creativity. This combination of business acumen and technical skill isn't easy to come by, and new graduate programs with an emphasis on data science seem to be cropping up daily to fill the gaps.

Open data science is proving to be a seedbed for innovation in the world economy. Open data science projects are revolutionizing the fabric of business in diverse industries, spawning new ecosystems of innovation.

Open team collaboration is essential for unlocking creativity in data science. Creativity comes when people from many backgrounds, roles and skill sets use open source data science tools—such as Apache Spark, R, and Apache Hadoop—to develop and deploy new designs for working and living. Data science initiatives foster innovation when teams combine the key roles and skill sets in pursuit of common objectives:

- Data scientists can use data science tools for teasing out the insights they're looking for and for making those insights actionable immediately through applications, visualizations and other consumables.
- Data engineers can build data processing pipelines that leverage machine learning, stream computing and other capabilities to ingest data from disparate sources, aggregate and cleanse it, and deliver it downstream to smart applications of all sorts.

SAITEJAXEROX

Scanned with CamScanner

- Business analysts can use statistical exploration tools to answer domain-specific questions quickly, easily and without the need of IT assistance.
- Application developers can use algorithmic capabilities to endow their applications with cognitive smarts that learn from fresh data and take actions that are continually optimized in keeping with contextual, predictive and environmental variables.

Tool capabilities

The pivotal importance of Spark and R in team data science stems from their ability, within open analytics ecosystems, to carry out several objectives:

- Facilitate the democratization of self-service data analytics development across enterprises and communities, especially when these programming tools are accessible from within teams' primary development workbench
- Enable distributed teams to address bigger data-centric problems and reap commensurately larger business results more rapidly than ever, especially when accessed in a shared, public cloud service
- Accelerate development of high-performance analytics applications rapidly, flexibly and easily, especially when using them with browser-based notebooks that support code, text, interactive visualization, math and media
- Provide a unified execution model for big data processing and analytics capabilities all in one environment, especially when deployed in conjunction with Apache Hadoop, NoSQL databases and other cloud-based data platforms
- Reduce the amount of code and number of tools needed to combine a deep stack of cognitive capabilities in a single application, especially when used in conjunction with rich libraries of machine learning, streaming analytics, graph computing, natural-language processing and other algorithms
- Allow teams to refine analytics applications interactively and iteratively, especially when used in conjunction with data and model governance features that are integrated into the data lakes around which the data science development lifecycle revolves

Tool innovation

Next-generation open analytics tools—such as IBM Data Science Experience (DSX)—provide a critical cloud-based, self-service enabler for decentralized teams to develop innovative data applications.

SAITEJA XEROX

Scanned with CamScanner